PHYSICSLENS: DIAGNOSING PHYSICAL PROPERTY BLINDNESS IN VIDEO GENERATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable video world models could provide scalable predictive environments for robot learning, planning, and evaluation. However, robot videos can violate physical principles and complete tasks through physically implausible behavior, limiting their reliability for robot learning and planning. Current video-generation benchmarks exclude physics that are inherently hidden by visuals (eg: weight, viscosity, friction). Due to this, video models are evaluated on the fidelity of physics-not the underlying accuracy of physics. We introduce PhysicsLENS, a dataset and benchmark for evaluating plausibility of physical properties grounded in robotics. PhysicsLENS uses matched scenario pairs that hold the same conditioning frame and task, while varying underlying physics in the scene description. Scenarios are curated from public robot video sources and annotated across seven physical domains: collision, gravity, momentum, friction, deformation, fluid, and causality. We evaluate across four video generation models, producing over 400 human-annotated labels. Results show a consistent accuracy gap between observable and unobservable conditions, indicating that models rely on visual appearance rather than grounded text-specified properties.

1 INTRODUCTION

Video models could provide scalable predictive environments for robot learning (Yang et al., 2024; NVIDIA, 2025a), planning (Du et al., 2023), and evaluation (Quevedo et al., 2025). By generating diverse interactions without repeated real-world data collection, they may expose robots to rare events, unfamiliar objects, and changing physical conditions (Jang et al., 2025). However, scaling the quantity of generated data is useful only if the resulting videos provide physically reliable supervision. A video may look realistic and show a robot completing its task while violating friction, momentum, object permanence, or other physical constraints. Such failures limit the value of generated videos as representations of how actions affect the physical world that are not measured in current benchmarks. We therefore need a benchmark that evaluates not only visual realism and task completion, but also whether generated interactions obey the physical properties that should govern them.

Existing video-generation benchmarks only partially measure this reliability. Most emphasize visual quality, temporal consistency, or the plausibility of motion visible in a generated video. These criteria can identify obvious artifacts, but they do not determine whether a model correctly grounds the physical properties specified in its conditioning input. Properties such as weight, friction, viscosity, and deformability may not be observable from a single image. Consequently, visually identical scenes can require different outcomes: a robot should interact differently with a heavy object than with a light one, or with a slippery surface than with a rough one. A model can therefore generate a plausible video while producing behavior that is physically incorrect for the stated condition.

We introduce PhysicsLENS, a benchmark for evaluating both physical plausibility and physical-property grounding in robot video generation. Its dataset contains curated image–text prompts, more than 400 clips generated by four video models, and structured human annotations for every clip. We construct prompts from public robot videos, grouping scenarios into eight physical domains: collision, gravity, contact, momentum, friction, deformation, fluid dynamics, and causality. Observable scenarios provide relevant physical cues in the initial frame; unobservable scenarios specify through text a property that cannot be determined from that frame. Matched unobservable prompts hold

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.

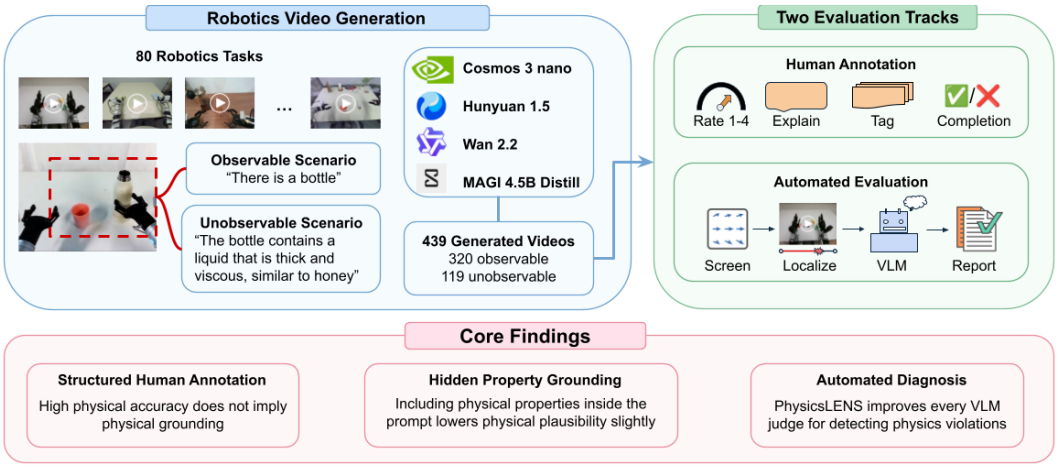


Figure 1: Overview of PhysicsLENS: robotics video generation, evaluation, and core findings.

the visual context and task fixed while varying the specified property, testing whether generated interactions reflect its physical consequences. Figure 1 summarizes the benchmark workflow.

The benchmark evaluates generated clips through two complementary tracks. First, human annotators rate each clip's physical plausibility and task completion, describe and label any visible violations, and, for unobservable scenarios, rate whether the stated property was followed. Second, a four-stage hierarchical pipeline screens each clip, localizes potential violations, verifies them with targeted vision-language model questions, and reports the supporting evidence. Across four video-generation models and 439 generated clips, the results show that a video's apparent physical plausibility does not guarantee correct grounding in a hidden property. Our contributions are:

1. We introduce PhysicsLENS, a robotics-focused benchmark covering eight physical domains and distinguishing physical plausibility from physical-property grounding.
2. We construct matched scenarios that preserve visual context while varying text-specified physical conditions, including properties that cannot be inferred from appearance alone.
3. We present a multi-stage agentic evaluation procedure and benchmark four video-generation models with more than 400 human judgments, revealing a consistent gap between observable and unobservable physical conditions.

2 Related Work

Evaluating physics in generated video. Video world models are increasingly used as simulators and data sources for robot learning (Yang et al., 2024; NVIDIA, 2025a; Jang et al., 2025), yet videos that look realistic can still break physics (Kang et al., 2025; Motamed et al., 2026). Distributional metrics and general benchmark suites mainly measure perceptual quality (Unterthiner et al., 2018; Huang et al., 2024), although VBench-2.0 adds physics dimensions (Zheng et al., 2025). Physics benchmarks score commonsense physics in generated videos (Bansal et al., 2025; 2026; Meng et al., 2025; Gu et al., 2026), compare generations with real experiments or recorded outcomes (Zhang et al., 2025; Motamed et al., 2026), or turn measured motion into verifiable rewards (Le et al., 2025). Physion-Eval and PQSG also localize failures, using expert reasoning traces and VLM-generated question graphs, respectively (Zhang et al., 2026; Pothiraj et al., 2026). For robotics, WorldSimBench, EWMBench, RBench, and RoboWM-Bench evaluate generated manipulation videos on visual quality, motion, task adherence, or whether the motion can be executed as actions (Qin et al., 2025; Yue et al., 2025; Deng et al., 2026; Jiang et al., 2026). WorldModelBench scores instruction following and physics adherence across domains that include robotics (Li et al., 2025), and RoboGaze reports temporally localized glitches under a robotics-specific taxonomy (Nguyen et al., 2026). To our knowledge, none of these benchmarks separates properties that are visible in the conditioning frame from properties stated only in the prompt (Table 1), so a high score cannot show whether a model followed the stated physics or only what the scene looks like.

Table 1: Comparison of physics benchmarks and video critics. Ref.-free requires no paired real-world recording; Obs. axis tests prompt-specified properties not inferable from the initial frame.

Category	Method	Physics	Robotics	Human	Ref.-free	Localized	Obs. axis
Physics benchmarks	VideoPhy-2 (Bansal et al., 2026)	✓	✗	✓	✓	✗	✗
	PhyGenBench (Meng et al., 2025)	✓	✗	✗	✓	✗	✗
	Morpheus (Zhang et al., 2025)	✓	✗	✗	✗	✗	✗
	Physics-IQ (Motamed et al., 2026)	✓	✗	✗	✗	✗	✗
	NewtonRewards (Le et al., 2025)	✓	✗	✗	✓	✗	✗
	GeoPhys (Internò et al., 2026)	✓	✗	✗	✓	✓	✗
Robotics evaluators	RoboGaze (Nguyen et al., 2026)	✓	✓	✓	✓	✓	✗
	RBench (Deng et al., 2026)	✓	✓	✓	✓	✗	✗
	RoboWM-Bench (Jiang et al., 2026)	✓	✓	✗	✗	✓	✗
Latent properties	Physion++ (Tung et al., 2023)	✓	✗	✓	✗	✗	✗
	T2V-CompBench (Sun et al., 2025)	✗	✗	✗	✓	✓	✗
VLM reasoning	Cosmos-Reason1 (NVIDIA, 2025b)	✓	✓	✗	✓	✓	✗
	QuantiPhy (Puyin et al., 2026)	✓	✗	✓	✓	✗	✗
	MASS (Wu et al., 2026)	✓	✗	✓	✓	✓	✗
This Work	**PhysicsLENS**	✓	✓	✓	✓	✓	✓

Hidden physical properties. Inferring properties that appearance does not reveal is a classic test of physical understanding: Physion++ and ComPhy ask models to infer mass, friction, elasticity, or charge from observed interactions (Tung et al., 2023; Chen et al., 2022). PhysicsLENS reverses this flow of information: the property is stated in text, and the generator must turn it into motion. Controllable generators receive physical quantities as dedicated inputs, such as forces or mass and friction coefficients (Liu et al., 2024; Gillman et al., 2025; Li et al., 2026). T2V-CompBench tests whether prompt attributes bind to the right objects, but only for visible attributes such as color, shape, and motion direction (Sun et al., 2025). We instead test whether general-purpose image-to-video models follow a property given only in plain language, with the frame and task held fixed.

VLM judges of physical plausibility. Automatic physics evaluation often relies on vision-language models (VLMs), either fine-tuned or prompted (Bansal et al., 2025; Meng et al., 2025; NVIDIA, 2025b), but their reliability is limited. VLMs struggle with physical-world understanding (Chow et al., 2025), most models score near chance on intuitive-physics tests (Bordes et al., 2025), and plausible verbal explanations can hide wrong estimates of velocity and acceleration (Puyin et al., 2026). Multimodal judges also show biases and inconsistent judgments (Chen et al., 2024), and video VLMs have limited fine-grained motion and temporal grounding (Hong et al., 2025; Wang et al., 2025). Measured evidence helps: MASS adds depth, object grounding, and motion tracking to VLM reasoning (Wu et al., 2026), and GeoPhys finds that VLM judges detect physics violations near chance while a training-free score on frozen image-encoder features does better (Internò et al., 2026). PhysicsLENS follows this direction by combining tracked measurements with targeted VLM questions, and we test it directly against human annotations.

3 PhysicsLENS Benchmark

3.1 Observable and Unobservable Conditions

Image-conditioned video generation relies on two sources of information: an initial frame and a text prompt describing the scene and requested task. In ordinary robot videos, visual appearance is often correlated with physical properties; for example, size may correlate with mass, and visible wetness may suggest low surface friction. A model may therefore generate plausible motion from visual priors without grounding the physical properties specified in text. PhysicsLENS evaluates whether models follow text-specified properties when those properties cannot be inferred from the initial frame. We define two conditions. In the *observable* condition, the prompt is consistent with the physical properties suggested by the frame. In the *unobservable* condition, the same frame is paired with a prompt that specifies a latent property that departs from those visual expectations. For example, a prompt may state that liquid inside an ordinary teapot is "thick and viscous, similar to molasses" (Figure 2). A model that relies only on appearance may instead generate water-like flow. We refer to this failure to condition on a text-specified latent property as *physical property blindness*.

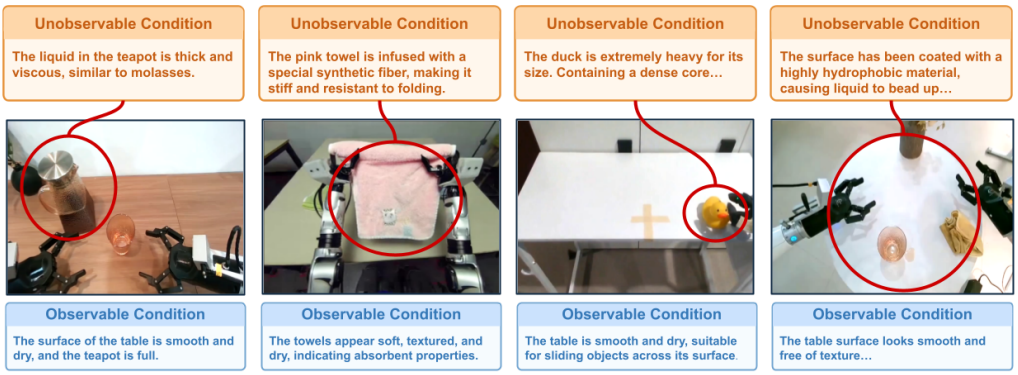


Figure 2: Example matched prompt pairs. Each pair shares the initial frame and task. The observable description (blue) states what the frame shows; the unobservable description (orange) replaces it with a latent property that departs from what the frame suggests.

Matched-pair construction. Each unobservable scenario is derived from an observable counterpart with the same initial frame and task instruction. Only the scene description changes: the observable description is consistent with the frame, whereas the unobservable description introduces a latent property that conflicts with its visual appearance. This construction holds the visual scene, task, and frame quality fixed, making the textual property the sole intervention. Because we generate only one video per prompt and model, individual pairwise differences may reflect generation noise; we therefore analyze matched pairs in aggregate. The unobservable scenarios specify four task-relevant properties: viscosity for pouring (10 scenarios), surface condition for wiping (8), elasticity for deformable objects (7), and mass for pushing (5). Each property is expressed in plain language within the scene description provided to the video generator. The generator receives neither a reference continuation nor the expected outcome.

The design supports two complementary evaluations. *Property grounding* assesses whether the generated video visibly reflects the stated latent property, using the human adherence score (Section 3.3). *Matched-condition comparison* tests whether specifying that property affects physical plausibility while holding the initial frame and task fixed—that is, whether failure to express the property coincides with visibly less plausible physics. Together, these evaluations distinguish general visual plausibility from adherence to the specified physical condition.

3.2 Scenarios and Generated Videos

PhysicsLENS contains 110 prompts: 80 observable scenarios and 30 matched unobservable variants. The 80 starting frames come from 11 public robot datasets that cover humanoids, dual-arm and single-arm manipulators, and mobile manipulators (Table 6). Each frame shows the scene just before a robot interacts with solid objects or liquids. Appendix A describes frame preparation, prompt generation and the annotation schema. We used GPT-4o to draft structured scene and action descriptions from each frame and its logged task. An author subsequently reviewed the descriptions.

We generate one video per prompt using each of four models: Wan 2.2 (Wan et al., 2025), Cosmos 3 Nano (Agarwal et al., 2026), HunyuanVideo 1.5 (Wu et al., 2025a), and MAGI 4.5B Distill (Teng et al., 2025). One unobservable HunyuanVideo 1.5 prompt did not produce a usable video. The realized benchmark hence contains 439 videos: 320 observable and 119 unobservable. HunyuanVideo 1.5 contributes 109 videos, including 29 unobservable videos, while each other model contributes 110. Since the benchmark contains one generation per prompt–model combination, the reported results characterize these generated clips and do not estimate variability across generation seeds.

3.3 Human Annotation

Annotators score each video on up to three scales, which we report separately because they can disagree. *Physical plausibility* $P \in \{1, 2, 3, 4\}$ judges only the visible physics: contact, motion, deformation, and object persistence. *Task completion* $A \in \{0, 1\}$ records whether the task is done

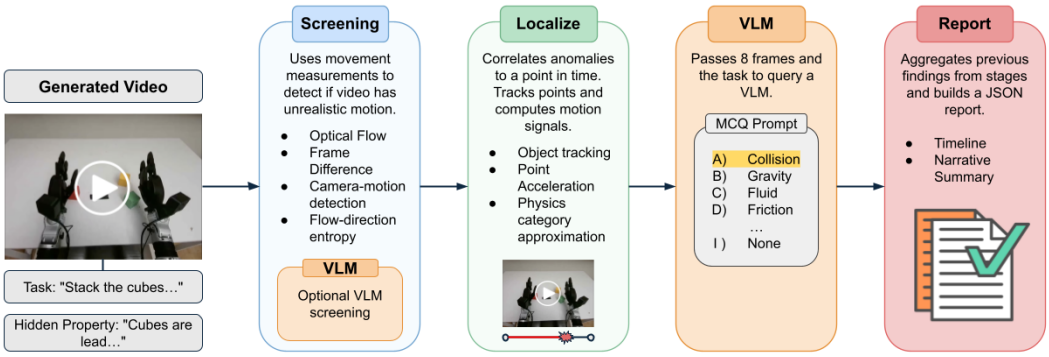


Figure 3: Evidence flow through the PhysicsLENS diagnostic pipeline.

by the end, even if the physics looks wrong. *Hidden-property adherence* $H \in \{1, 2, 3, 4\}$, rated only for unobservable videos, records whether the outcome shows the stated property. We treat $P \geq 3$ as mostly or fully plausible and $H \geq 3$ as mostly or fully following the property.

When $P < 4$, annotators must describe the physics violation in the video, and select one or more of eight violation labels. These labels include: collision, contact, friction, momentum, fluid, deformation, gravity, and causality. Appendix B gives the full rubric, the label definitions, and how videos were assigned to annotators.

3.4 Automated Diagnostic Pipeline

PhysicsLENS also includes an automated evaluator (Figure 3). Because VLM judges are unreliable on their own (Section 2), the evaluator pairs motion measurements, which record how objects and the camera move, with targeted VLM questions, which identify what went wrong. Its input is the generated video and the task. For unobservable scenarios, the hidden-property question also receives the stated property and the scenario's expected outcome. Human labels are used only to fit how the evidence is combined, on training folds that exclude the scenario being judged.

The evaluator has four stages. *Screening* measures motion over the first 5 seconds of the video: frame differences, optical flow, camera motion, flow-direction entropy, and how DINOv2 frame embeddings change. It also asks a VLM a physics-error question that lists specific errors instead of asking about overall realism. *Localization* tracks points with camera motion removed and measures their kinematics: speed, acceleration, the time of peak acceleration, and points that disappear inside the frame rather than at its border, the sign of an object vanishing mid-video. *VLM verification* shows each VLM eight frames and the task and asks three questions: (i) whether the task was completed; (ii) which of eight failure descriptions (one per constraint family, plus an object appearing or vanishing) best describes the main physics problem, if any; and (iii) for unobservable scenarios, whether the outcome shows the stated property. *Reporting* merges the evidence into a per-video JSON report with each finding's time window, object, violation type, confidence, and a short explanation. We do not evaluate time localization because our human labels are not time-dependent.

To compare with human ratings (Section 4.4), we pose these questions to ten VLMs and read each answer from the probabilities of its first output token. For each question, we rank each VLM's scores across videos, average the ranks over the ten VLMs, and mix the result with five screening and localization signals using a single weight. The five signals and the weight are selected on the training folds. Reports are not scored, because the human labels have no time stamps.

4 Experimental Results

This section presents results on the PhysicsLENS benchmark and evaluates its diagnostic pipeline through four research questions. RQ1–RQ3 are answered with human annotations, and RQ4 compares the pipeline against them. **RQ1:** Does high physical plausibility imply that a model followed a grounded specification (Sec. 4.1)? **RQ2:** Does stating a latent property in text change plausibility (Sec. 4.2)? **RQ3:** What do individual failure cases reveal about the distinction between grounding

Table 2: Human evaluation on unobservable scenarios. $\bar{P}$, $\bar{H}$: mean plausibility and adherence (1–4). Compl.: completion rate.

Model	n	$\bar{P}$	$\bar{H}$	Compl.	$H{\leq}2 \mid P{\geq}3$
HunyuanVideo 1.5	29	2.24	2.24	72.4%	5/13 (38.5%)
MAGI 4.5B distill	30	3.10	1.10	6.7%	21/21 (100%)
Wan 2.2	30	1.53	1.90	40.0%	2/5 (40.0%)
Cosmos 3 Nano	30	2.07	1.87	66.7%	6/8 (75.0%)

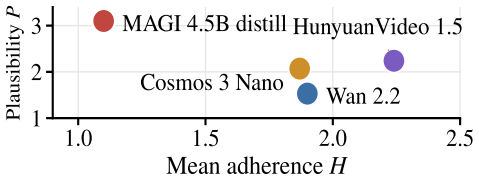


Figure 4: Mean plausibility (P) vs. mean adherence (H) per model (1–4).

and plausibility (Sec. 4.3)? **RQ4:** Can the PhysicsLENS pipeline approximate human judgments (Sec. 4.4)?

4.1 RQ1: Does High Plausibility Imply Property Grounding?

A video may look physically plausible without reflecting the physical property specified in its prompt. To test this, we collected annotations rating each video's plausibility (P) and its adherence to the stated property (H) on 1–4 scales (Sec. 3.3). We consider a video plausible when $P \geq 3$ and property-adherent when $H \geq 3$.

Across the total of 119 videos with unobservable properties, human annotators noted 47 as plausible. Yet 34 of those 47 (72%) do not adhere to the stated property ($H \leq 2$). This is the *physical property blindness* described in Section 3.1. The pattern appears for all four generators (Table 2). MAGI has the highest mean plausibility but the lowest mean adherence (Fig. 4); it accounts for 21 of the 34 plausible but ungrounded videos. The result persists without MAGI: 13 of the remaining 26 plausible videos (50%) ignore the property.

On the other hand, 28 of the 119 videos that adhere to the stated property ($H \geq 3$), 15 are rated implausible ($P \leq 2$). These videos reflect the implied property but violate other aspects of physics. **Thus, plausibility and property grounding must be evaluated separately: a video can satisfy either criterion without satisfying the other.**

4.2 RQ2: Does Stating A Latent Property In Text Change Plausibility??

We next compare how the same physical property is presented. Each unobservable scenario has an observable counterpart with the same scene, action, and generator. Each unobservable scenario has an observable counterpart with the same frame, task, and generator. Only the description differs: it either matches the frame or specifies a different latent property (e.g., molasses-like tea in an ordinary teapot). This comparison tests whether the presentation of the property changes the plausibility of the generated video.

Mean plausibility is lower in the unobservable condition for all four generators, by 0.13 to 0.38 points on the 1–4 scale (Table 3). Across individual pairs, plausibility decreases in 40 cases and increases in 26. Pooling the 29 scenarios generated by all four models gives a mean decrease of 0.28 points (Fig. 5; $p = 0.085$). **Thus plausibility for current video models tends to show no correlation with how properties are stated in text.**

4.3 RQ3: What Do Failure Cases Reveal About Grounding and Plausibility?

Fig. 6 illustrates why a plausibility rating alone cannot establish property grounding. In a task that specifies molasses-like tea, all four generators produce plausible pours that fail to reflect its viscosity ($H{=}1$). Wan 2.2, for example, pours the tea like water while receiving the highest plausibility rating (Fig. 6a). A different case arises when the requested interaction barely occurs. MAGI's robot often remains nearly still (Fig. 6c), producing a video that looks plausible without demonstrating how the stated property affects the task.

On the contrary, a video can reflect the stated property while violating other physical constraints. With a silicone-coated table, HunyuanVideo 1.5 receives an adherence rating of $H{=}3$, but the glass moves without contact, making the video implausible (Fig. 6b). In another case, Wan 2.2 leaves an extremely heavy tape dispenser stationary ($H{=}4$), but the robot's hand passes through it ($P{=}1$).

Table 3: Matched plausibility comparison. Obs., Unobs.: mean P on the matched pairs; Δ: unobservable minus observable. p: Holm-adjusted paired test; pooled row: scenario-level sign-flip test over the 29 scenarios annotated for all four models.

Model	n	Obs.	Unobs.	Δ	p
HunyuanVideo 1.5	29	2.62	2.24	−0.38	0.709
MAGI 4.5B distill	30	3.33	3.10	−0.23	0.884
Wan 2.2	30	1.83	1.53	−0.30	0.497
Cosmos 3 Nano	30	2.20	2.07	−0.13	0.884
Pooled	29	—	—	−0.28	0.085

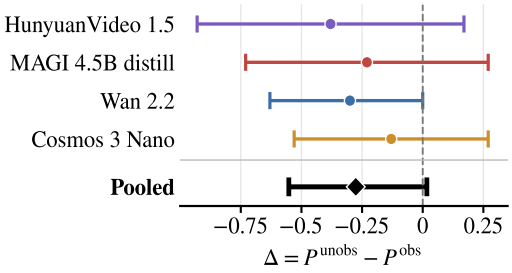


Figure 5: Paired change $\Delta = P^{\text{unobs}} - P^{\text{obs}}$ per model and pooled, with bootstrap intervals over scenarios.

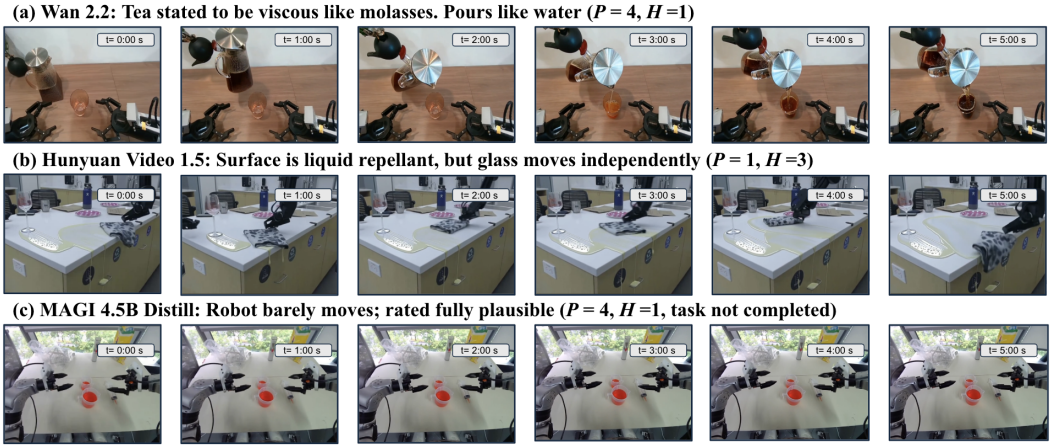


Figure 6: Three annotated failure cases (six frames each). P: physical plausibility; H: hidden-property adherence (both 1–4). (a) A clean, realistic pour that ignores the stated viscosity. (b) A video rated as following the stated property that is nevertheless physically implausible. (c) A nearly static video rated fully plausible.

PhysicsLENS ranks the visible violation in Fig. 6b among the benchmark's least plausible videos. **These cases show why physical plausibility and adherence to the stated property require separate evaluations.**

4.4 RQ4: How Well Do Automated Evaluators Match Human Judgments on PhysicsLENS?

PhysicsLENS provides human annotations for 439 videos, allowing us to measure how well automated scores agree with human judgments. We evaluate ten open-weight VLMs from seven families (7B–32B parameters), Stage-1/2 video signals, and combinations of the two. Combination weights and signal selection are fitted using five-fold cross-validation, with all versions of a scenario kept in the same fold. We report AUC, where 0.5 is chance and 1 is perfect ranking.

Combining VLM judgments with video signals improves agreement. The full evaluator combines answers from an ensemble of ten VLMs with Stage-1/2 signals. It reaches an AUC of 0.72 for overall plausibility, 0.81 for task completion, 0.73 for violation detection, and 0.71 for hidden-property adherence (Table 4). It achieves the highest or tied-highest score on five of the seven measures. Adding signals to an individual VLM raises mean overall plausibility AUC from 0.53 to 0.61, violation detection from 0.58 to 0.72, task completion from 0.71 to 0.76, and property adherence from 0.57 to 0.70. Signals alone reach 0.70 for violation detection; one frequently selected signal tracks objects vanishing during a video.

Generator identity is a strong baseline because the four generators differ substantially in quality. It slightly exceeds the full evaluator on unobservable-video plausibility (0.72 versus 0.71), but assigns

Table 4: PhysicsLENS against human judgment (AUC on 439 videos; 0.5 is chance). *Generator identity* scores each video by its generator's average rating; *signals* are the PhysicsLENS Stage-1/2 measurements; *debiased* is the PhysicsLENS physics-error question; the *ensemble* averages all ten VLMs. Values are mean ± standard deviation across the three runs, except for † rows, which report mean ± standard deviation across the ten VLMs. Bold indicates the best result in each column.

	Physical plausibility			Task	Violation	Physics	Hidden
System	Obs.	Unobs.	All	completion	detection	family	property
Generator identity	0.63±0.03	**0.72±0.05**	0.65±0.03	0.72±0.02	**0.73±0.03**	0.51±0.02	0.68±0.05
Signals only	0.61±0.03	0.65±0.05	0.62±0.03	0.70±0.03	0.70±0.03	0.51±0.02	0.69±0.05
VLM†	0.53±0.06	0.53±0.06	0.53±0.05	0.71±0.04	0.58±0.08	0.59±0.02	0.57±0.03
VLM + signals†	0.60±0.02	0.66±0.01	0.61±0.01	0.76±0.02	0.72±0.01	0.58±0.02	0.70±0.02
Debiased VLM†	0.64±0.07	0.63±0.04	0.63±0.06	0.71±0.04	0.58±0.08	0.59±0.02	0.57±0.03
Debiased + signals†	0.65±0.05	0.67±0.03	0.65±0.04	0.76±0.02	0.72±0.01	0.58±0.02	0.70±0.02
PhysicsLENS ensemble	**0.73±0.03**	0.71±0.05	**0.72±0.02**	0.77±0.02	0.67±0.03	**0.63±0.01**	0.59±0.06
PhysicsLENS (full)	**0.73±0.03**	0.71±0.05	**0.72±0.02**	**0.81±0.02**	**0.73±0.03**	0.62±0.02	**0.71±0.05**

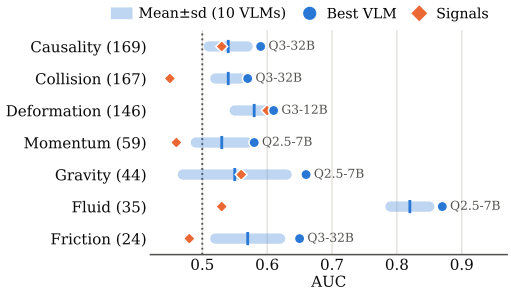


Figure 7: Attribution accuracy by physics family: AUC for videos with this family vs. videos with a different violation (dotted line is chance).

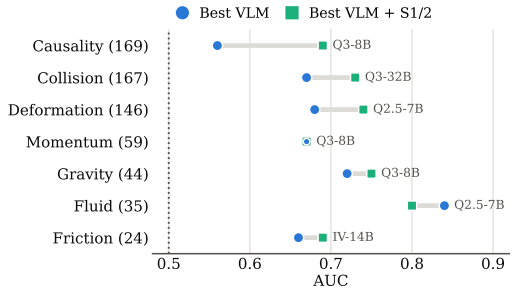


Figure 8: Detection accuracy by physics family: AUC for videos with this family vs. the 106 videos with no violation (dotted line is chance).

the same score to every video from a given generator. Its within-generator AUC is therefore 0.5 by construction, whereas the automated evaluator reaches 0.67 for unobservable plausibility and 0.76 for task completion. Generator identity also cannot score videos from a new generator without first obtaining human ratings for that generator.

Targeted questions improve VLM plausibility judgments. When asked directly whether a video is physically plausible, many VLMs perform near chance (Table 5). To focus their judgments on physical behavior, we instead ask about specific errors, such as objects passing through one another, vanishing, floating, changing shape impossibly, or moving without contact. The question also instructs the VLMs to disregard task success and the amount of motion. It improves plausibility AUC for eight of the ten VLMs. On observable videos, Qwen3-VL-32B improves from 0.55 to 0.75 and InternVL3-14B from 0.51 to 0.71. On VideoPhy-2 (Bansal et al., 2026), Qwen3-VL-32B improves from 0.645 to 0.720.

Hidden-property adherence remains difficult to judge. With the targeted physics-error question, Qwen3-VL-32B reaches plausibility AUCs of 0.75 on observable videos and 0.76 on unobservable videos. Its AUC for judging whether a text-specified property was followed, however, is only 0.57. Across the ten VLMs, adding signals raises mean adherence AUC from 0.57 to 0.70, close to signals alone (0.69) and generator identity (0.68). This agreement does not establish that the automated evaluator can determine whether a particular hidden property was realized.

Violation detection is easier than attribution. The best VLM's detection AUC ranges from 0.56 to 0.84 across seven violation families; adding video signals improves five families, leaves momentum unchanged, and lowers performance for fluids (Figure 8). Among videos with different violations, mean VLM attribution AUC ranges from 0.53 to 0.58 for most families, although fluids reach 0.82 (Figure 7). Since 66% of videos with violations contain multiple types, assigning a single fam-

Table 5: Plausibility AUC per VLM on observable (n=320) and unobservable (n=119) videos, with a standard question, the PhysicsLENS physics-error question, and that question plus Stage-1/2 signals. *Hidden property*: whether the stated property was followed. Mean ± sd over three runs that show the VLM different frames of each video. Bold: best per column.

	Standard		Physics-error		+ signals		Hidden
VLM	Obs.	Unobs.	Obs.	Unobs.	Obs.	Unobs.	property
Qwen3-VL-32B	0.55±0.00	0.57±0.02	**0.75±0.00**	**0.76±0.05**	**0.74±0.00**	**0.76±0.05**	0.57±0.01
InternVL3-14B	0.51±0.01	0.45±0.02	0.71±0.01	0.65±0.00	0.71±0.02	0.68±0.01	0.51±0.01
Qwen3-VL-8B	0.54±0.01	0.58±0.01	0.68±0.00	0.70±0.01	0.68±0.01	0.71±0.01	0.56±0.01
Qwen2.5-VL-32B	0.45±0.01	0.44±0.01	0.66±0.02	0.64±0.02	0.67±0.01	0.71±0.01	0.57±0.01
InternVL3-8B	**0.63±0.00**	**0.62±0.02**	0.65±0.01	0.59±0.01	0.65±0.01	0.62±0.01	0.60±0.01
Llama-4-Scout-17B	0.57±0.01	0.50±0.02	0.63±0.02	0.59±0.04	0.63±0.01	0.65±0.01	0.60±0.02
Qwen2.5-VL-7B	0.46±0.02	0.53±0.01	0.60±0.02	0.61±0.03	0.61±0.02	0.66±0.01	0.56±0.01
Gemma-3-12B	0.61±0.01	**0.62±0.02**	0.59±0.00	0.60±0.01	0.60±0.00	0.65±0.01	0.59±0.03
LLaVA-OV-7B	0.47±0.00	0.55±0.01	0.55±0.01	0.62±0.00	0.59±0.00	0.66±0.01	0.62±0.00
Mistral-Small-3.1-24B	0.53±0.01	0.54±0.01	0.53±0.00	0.58±0.02	0.58±0.01	0.66±0.01	0.56±0.01
Signals only	—	—	—	—	0.60±0.01	0.65±0.00	**0.69±0.05**

ily is particularly difficult. **Automated scores agree with human judgments on several visible outcomes, but hidden-property grounding and precise violation attribution remain difficult.**

5 Conclusion

We introduced PhysicsLENS, a benchmark for testing whether generated robot videos actually reflect the physical properties they're told to have. By keeping the initial frame and action the same and only changing how the property is described, matched scenarios let us look at property grounding and physical plausibility side by side. Across 439 clips from four generators, human evaluation found a clear gap between the two: 34 of 47 clips rated plausible under unobservable conditions didn't follow the stated property, and 15 of 28 clips that did follow it were still physically implausible. In other words, a video can look right without actually respecting the physics it was asked to show, or get the property right while breaking something else entirely. Pairing targeted VLM questions with measured video signals meaningfully improves agreement with these human judgments, reaching an AUC of 0.72 for plausibility and 0.81 for task completion, though judging whether a specific hidden property was followed, and figuring out which physics family a violation belongs to, are both still hard problems. Taken together, these results make the case for evaluating plausibility and property grounding as two separate questions, not one.

We kept this study deliberately focused to 80 initial frames and four video-generation models so we could isolate property grounding under tightly matched conditions. From here, a natural next step is testing more models, more hidden-property types, and multiple generations per prompt, which would help separate how much of the variation comes from the generation process itself versus the property being tested.

AI USE STATEMENT

In this work, we used generative AI tools for: drafting portions of the final text, exploring up-to-date literature, and generating code for the diagnostic tool and annotation pipeline. We have not used generative AI tools for: generating experimental results or designing core methodology. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aligns with the ICLR Code of Ethics. All datasets used are publicly available and licensed for research use. Annotations were produced by members of the research team rather than externally recruited participants, and did not involve the collection of personal, sensitive, or identifying information. We affirmed that all results and methods are explained accurately, transparently, and reproducibly. This work aligns with responsible and beneficial use of video generation research for robotics.

REPRODUCIBILITY STATEMENT

For details on how to use our data and evaluation, check Appendix C.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

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. In *ICLR*, 2026.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, et al. LeRobot: State-of-the-art machine learning for real-world robotics in PyTorch. `https://github.com/huggingface/lerobot`, 2024.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR, 2024.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. ComPhy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations*, 2022.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world, 2026. URL https://arxiv.org/abs/2601.15282.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660, 2024.

Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4066–4083. PMLR, 2025.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. DreamDojo: A generalist robot world model from large-scale human videos. In *International Conference on Machine Learning (ICML)*, 2026.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. "phyworldbench": A comprehensive evaluation of physical realism in text-to-video models, 2026. URL https://arxiv.org/abs/2507.13428.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8450–8460, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Christian Internò, Alexander Pondaven, Habon Issa, Fabio Pizzati, Francesco Pinto, Markus Olhofer, Ivan Laptev, Philip Torr, Eero P. Simoncelli, Barbara Hammer, and David Klindt. Geophys: The geometry of physical plausibility, 2026. URL https://arxiv.org/abs/2606.20707.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. DreamGen: Unlocking generalization in robot learning through video world models. In *Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5170–5194. PMLR, 2025.

Feng Jiang, Yang Chen, Kyle Xu, Yuchen Liu, Haifeng Wang, Zhenhao Shen, Jasper Lu, Shengze Huang, Yuanfei Wang, Chen Xie, and Ruihai Wu. Robowm-bench: A benchmark for evaluating world models in robotic manipulation, 2026. URL https://arxiv.org/abs/2604.19092.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28991–29017. PMLR, 2025. URL https://proceedings.mlr.press/v267/kang25g.html.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models, 2025. URL https://arxiv.org/abs/2502.20694.

Yanxun Li, Hao Wen, Bingze Song, Jiashu Zhu, Aiming Hao, Chubin Chen, Jintao Chen, Jiahong Wu, Xiangxiang Chu, and Miao Wang. Learning explicit physical parameter control and benchmarking for video generation. *arXiv preprint arXiv:2607.18924*, 2026.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, volume 15140 of *Lecture Notes in Computer Science*, pp. 360–378. Springer, 2024.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *ICML*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *WACV*, pp. 948–958, 2026.

Minh-Loi Nguyen, Nghiem Tuong Diep, Hung Khang Nguyen, Minh Le, Doanh Le Thien, Hoang H Tran, Dung D Le, Vu N Duong, Daniel Sonntag, An Thai Le, et al. Robogaze: Evaluating robot world models via structured vision-language analysis. *arXiv preprint arXiv:2606.28385*, 2026.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025a.

NVIDIA. Cosmos-Reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025b.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Abhishek Padalkar, Gabriel Quere, Franz Steinmetz, Antonin Raffin, Matthias Nieuwenhuisen, João Silvério, and Freek Stulp. Guiding reinforcement learning with shared control templates. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. doi: 10.1109/ICRA48891.2023.10161058.

Atin Pothiraj, Jaemin Cho, Yue Zhang, Elias Stengel-Eskin, and Mohit Bansal. Physics question scene graph: Fine-grained evaluation of physical plausibility in text-to-video generation, 2026. URL https://arxiv.org/abs/2606.25306.

Li Puyin, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33174–33184, 2026.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, LEI BAI, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=j9pVnmulQm.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation, 2025. URL https://arxiv.org/abs/2407.14505.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel LK Yamins, Judith E Fan, and Kevin A. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties, 2023. URL https://arxiv.org/abs/2306.15668.

Unitree Robotics. G1_Dex1 datasets. https://huggingface.co/datasets/unitreerobotics/G1_Dex1_MountCamera_Dataset, 2025a. Hugging Face dataset repositories unitreerobotics/G1_Dex1_*.

Unitree Robotics. G1_Dex3 datasets. https://huggingface.co/datasets/unitreerobotics/G1_Dex3_BlockStacking_Dataset, 2025b. Hugging Face dataset repositories unitreerobotics/G1_Dex3_*.

Unitree Robotics. Z1_Dual_Dex1_PourCoffee_Dataset. https://huggingface.co/datasets/unitreerobotics/Z1_Dual_Dex1_PourCoffee_Dataset, 2025c.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 959–975, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025a.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. RoboCOIN: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025b.

Xiyang Wu, Zongxia Li, Jihui Jin, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. Mass: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9433–9443, 2026.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024.

Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

Chenyu Zhang, Daniil Cherniavskii, et al. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-eval: Evaluating physical realism in generated video via human reasoning, 2026. URL https://arxiv.org/abs/2603.19607.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. doi: 10.15607/RSS.2023.XIX.016.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

A Dataset Details

A.1 Source Datasets and Licenses

The 80 starting frames come from demonstrations in the 11 public robot datasets listed in Table 6. For each scenario we use only the first frame of one demonstration. Most sources are LeRobot-format releases (Cadene et al., 2024). Licenses are those stated by each source. AgiBot World, GR00T-Teleop-GR1, and part of RH20T carry non-commercial or share-alike terms, and the UniBot-V1 challenge repositories in the Unitree G1 Dex1 collection state no license. We release each frame together with its source and license and do not relicense any frame.

Table 6: Public datasets that supply the starting frames. Source is the Hugging Face repository we used; † marks an unofficial copy, with the official release in parentheses.

Dataset	Robot (type)	License	Source
Humanoid Everyday (Zhao et al., 2026)	Unitree G1, H1 (humanoid)	Apache-2.0	`USC-GVL/humanoid-everyday`
RoboCOIN (Wu et al., 2025b)	AlphaBot 2, Galbot G1, Unitree G1edu (dual-arm, humanoid)	Apache-2.0, with use terms	`RoboCOIN/*`
Unitree G1 Dex1 (Unitree Robotics, 2025a)	Unitree G1 with Dex1 grippers (humanoid)	Apache-2.0; none stated for UniBot-V1 challenge sets	`unitreerobotics/G1_Dex1_*`
Unitree G1 Dex3 (Unitree Robotics, 2025b)	Unitree G1 with Dex3 hands (humanoid)	Apache-2.0	`unitreerobotics/G1_Dex3_*`
Unitree Z1 dual-arm (Unitree Robotics, 2025c)	Two Unitree Z1 arms (dual-arm)	Apache-2.0	`unitreerobotics/Z1_Dual_Dex1_PourCoffee_Dataset`
GR00T-Teleop-GR1 (Gao et al., 2026)	Fourier GR-1 (humanoid)	CC BY-NC 4.0	`nvidia/PhysicalAI-Robotics-GR00T-Teleop-GR1`
AgiBot World Alpha (Bu et al., 2025)	AgiBot G1 (dual-arm, mobile)	CC BY-NC-SA 4.0	`weijian-sun/agibotworld-lerobot`† (`agibot-world/AgiBotWorld-Alpha`)
Mobile ALOHA (Fu et al., 2025)	Mobile ALOHA (dual-arm, mobile)	MIT or Apache-2.0, by repository	`lerobot/aloha_mobile_*`
ALOHA (Zhao et al., 2023)	Static ALOHA (dual-arm)	MIT or Apache-2.0, by repository	`lerobot/aloha_static_*`
RH20T (Fang et al., 2024)	Flexiv arm (single-arm)	CC BY-SA 4.0 or CC BY-NC 4.0, by scene	`robot-lev/rh20t_cfg1`† (`rh20t.github.io`)
DLR SARA Pour (Padalkar et al., 2023)	DLR SARA (single-arm)	CC BY 4.0	`lerobot/dlr_sara_pour`, from Open X-Embodiment (Open X-Embodiment Collaboration et al., 2024)

A.2 Image and Prompt Preparation

We selected a conditioning frame immediately before the target manipulation and standardized the supplied frames to 1280×720 pixels. GPT-4o received each frame and its logged task under a fixed two-field JSON instruction. It produced a description of the visible scene and a one-sentence description of the complete manipulation without specifying its outcome. We reviewed these descriptions and assembled the generation prompt from the `Scene` and `Action` fields. For unobservable variants, we retained the same frame and action while rewriting the scene description to state the hidden property. Expected outcomes and failure signatures were used for evaluation and excluded from the generation prompt.

A.3 Taxonomy and Curation Notes

Collision and Friction scenarios were comparatively easy to source, reflecting the prevalence of pick-and-place and surface-manipulation footage. Fluid and Deformation scenarios were harder to find, while Gravity scenarios that isolated free fall were rare outside incidental object drops. Figure 10 reports the resulting domain distribution. These observations describe our curation process only; establishing the distribution of public robotics video would require a systematic survey of available sources.

Annotation procedure. Scenarios were curated from public robotics video sources rather than authored domain-first, and each scenario was subsequently annotated with its governing physical domain by manual review of its action and outcome. This annotation was performed independently

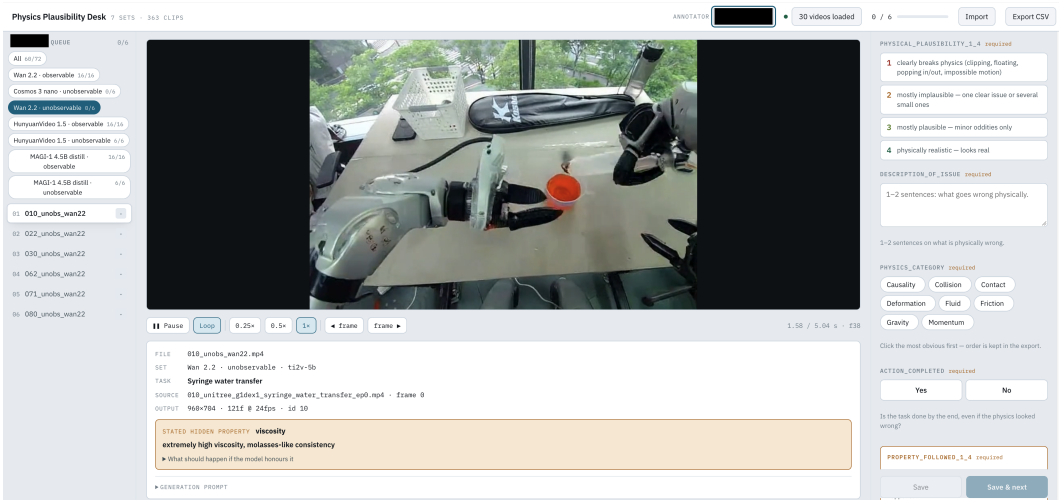


Figure 9: The human-annotation interface. Annotators assess physical plausibility and action completion for every generated clip, identify visible violations when present, and evaluate property adherence for unobservable scenarios.

of the taxonomy's construction to avoid circularity: the seven domains were fixed in advance from the specialist families, and each scenario was assigned to the domain its physics most directly tests, rather than domains being defined post hoc to fit whatever scenarios happened to be available.

Interaction with observability. The observable/unobservable axis introduced in §3.1 is orthogonal to domain: a scenario's domain label is fixed by the physical law under test, while its observability condition determines whether the property governing that law is visually apparent or text-specified. Both matched members of an observable/unobservable pair therefore carry the same domain label, allowing the accuracy gap described in §3.1 to be reported per domain rather than only in aggregate.

B Human Annotation

B.1 Annotation Instrument

The annotation interface presents one generated clip at a time together with the fields summarized in Table 7. Physical plausibility and action completion are collected for every clip. Hidden-property adherence is collected only for unobservable scenarios, for which the form includes a property-specific question. When a clip receives a plausibility score below 4, annotators must also describe the issue and select at least one violation category. Figure 9 shows the interface presented to annotators.

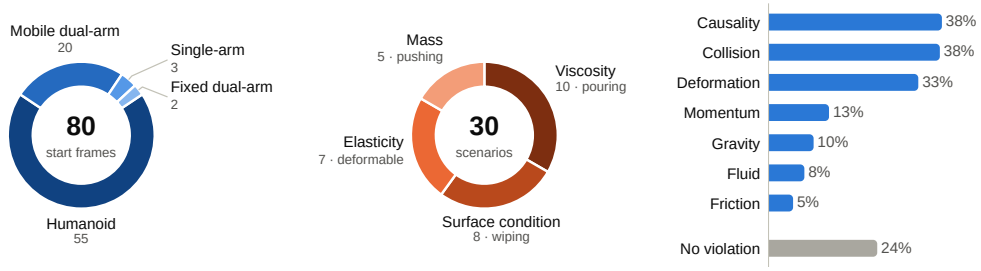
Interpretation of the scores. The physical-plausibility score $P \in 1, 2, 3, 4$ assesses only the visible consistency of contact, motion, deformation, and object persistence. It is a human judgment rather than a measurement against a physical simulator. Task incompletion alone does not imply physical implausibility.

The action-completion score $A \in 0, 1$ assesses the final task outcome rather than the physical correctness of the process. A clip may therefore receive $A = 1$ while exhibiting severe physical violations, or $A = 0$ despite otherwise plausible motion.

The hidden-property score $H \in 1, 2, 3, 4$ measures whether the generated behavior demonstrates the property stated in the prompt. It is not collected for observable scenarios, and missing values in those scenarios are not interpreted as zero or as failures. Importantly, a low adherence score does not distinguish between two cases: the video may visibly contradict the stated property, or it may fail to perform an interaction that would reveal the property. Because the instrument does not provide a separate "insufficient evidence" response, we retain this ambiguity when interpreting $H \leq 2$.

Table 7: Fields in the human-annotation form. Monospace text gives each field's identifier. P: physical plausibility; A: action completion; H: hidden-property adherence.

Field	Response	Collected for	Instructions
Physical plausibility (P) `physical_plausibility`	1–4	All clips	**1**: clearly breaks physics, including clipping, floating, objects appearing or disappearing, or impossible motion. **2**: mostly implausible, with one clear issue or several smaller issues. **3**: mostly plausible, with minor oddities only. **4**: physically realistic.
Issue description `description_of_issue`	Free text, 1–2 sentences	Clips with $P < 4$	Briefly describe what is physically incorrect. Leave blank when $P = 4$.
Violation labels `physics_category`	Multi-select, 8 labels	Clips with $P < 4$	Select at least one applicable violation category, choosing the most obvious issue first. Leave blank when $P = 4$.
Action completion (A) `action_completed`	Yes/no	All clips	Judge whether the requested task is completed by the end of the clip, even if the process is physically unrealistic.
Hidden-property adherence (H) `property_followed`	1–4	Unobservable clips	**1**: property completely ignored; the outcome resembles a generic response. **2**: mostly ignored, with only a faint or inconsistent indication. **3**: mostly followed, but with incorrect magnitude, timing, or degree. **4**: fully and correctly reflected.



(a) Robot type (b) Hidden property (c) Violation label

Figure 10: Dataset composition. (a) Robot type in the 80 starting frames. Mobile dual-arm robots are wheeled platforms (AlphaBot 2, Galbot G1, AgiBot G1, and Mobile ALOHA). (b) Property stated in each unobservable scenario and the task it is paired with. (c) Share of the 439 videos with each violation label. As in Appendix C, Contact is merged into Collision. A video can have several labels (66% of videos with a violation have more than one), so the violation bars sum to over 100%.

Aggregation. We report rating distributions, descriptive means of the ordinal scores, and action-completion proportions. For thresholded summaries, $P \geq 3$ denotes mostly plausible or physically realistic behavior, while $H \geq 3$ denotes that the stated property is mostly or fully followed. We additionally examine the joint distribution of P and H, including the proportion of apparently plausible clips ($P \geq 3$) that do not follow the hidden property ($H \leq 2$). The three dimensions are reported separately rather than combined into a single quality score.

B.2 Violation-Label Codebook

The `physics_category` field contains the eight labels defined in Table 8. It is shown for every clip and requires at least one selection when $P < 4$. Annotators are instructed to select the most

obvious violation first. Although click order is preserved in the exported data, our frequency analyses treat the selected labels as an unordered set; order is not interpreted as a calibrated ranking of severity.

Table 8: Definitions of the eight violation labels used in the human annotations.

Label	Definition
Causality	An effect occurs without a visible cause, or events occur in an impossible order.
Collision	Objects pass through one another or fail to exhibit an expected response after impact.
Contact	An object floats without support, rests at an incorrect contact point, or fails to maintain visible contact.
Deformation	A rigid object warps or melts, or a soft object fails to deform under an interaction that should alter its shape.
Fluid	Pouring, splashing, containment, or other liquid behavior is physically inconsistent.
Friction	Sliding, sticking, or gripping behavior is inconsistent with the visible interaction or stated surface condition.
Gravity	An object floats, falls incorrectly, or otherwise fails to respond plausibly to gravity.
Momentum	An object accelerates or changes motion without a supporting interaction, including apparent creation or loss of energy.

These eight human labels are distinct from the seven physical families used to organize the benchmark and the automated specialists. In particular, the human codebook includes both **Contact** and **Collision**, whereas the specialist taxonomy contains seven families. We preserve the human labels as recorded and merged Contact into Collision. A blank category field indicates only that no label was recorded; it is not independent confirmation that the clip is physically correct.

Some violations may reasonably receive more than one label. For example, an object passing through a supporting surface may involve both Contact and Collision, while unsupported floating may involve both Contact and Gravity. The multi-select interface preserves these overlaps rather than forcing each failure into a single category.

B.3 Annotator Allocation and Reliability Limits

Annotators were recruited from a computer science research environment and included mainly graduate students. Clip assignment is random. All observable clips were distributed in batches of 16–17 clips per annotator. Unobservable clips were distributed in batches of 6.

Some observable Cosmos records contain a combined annotator label. Because these records contain only one final set of scores rather than separate ratings attributable to each individual, we treat the combined label as a single annotation source. It does not constitute overlapping independent annotation and cannot be used to estimate agreement between the named annotators. Each clip received one final annotation rather than independent ratings from multiple annotators. Consequently, the dataset does not support an estimate of inter-annotator reliability.

C PhysicsLENS Pipeline Details

This appendix describes the pipeline as it was run for Section 4.4. Where the interactive tool of Section 3.4 does more, we say so.

Screening. Stage 1 measures motion over the first 5 s of each video, sampled at 8 frames per second in grayscale with the longer side resized to 320 pixels. It computes the mean absolute difference between consecutive frames; dense optical flow (Farnebäck); camera motion, taken as the median flow vector of each frame pair; object motion, the flow left after subtracting camera motion; and the entropy of object-motion directions over the fastest-moving 10% of pixels. It also embeds the eight VLM frames (below) with DINOv2 and measures how the embedding moves: the step size between frames, how that step size changes, whether successive steps keep their direction, and net displacement relative to path length. Each measurement is reduced to one number per video (a mean, 95th percentile, maximum, or ratio), giving 22 screening signals. No signal is thresholded

into a flag. Signals become scores only when evidence is combined, by ranking videos against each other (below). Stage 1 also asks the VLM the physics-error question (Table 9). It asks the model to count specific physics errors (objects passing through each other, appearing or vanishing, floating, changing shape impossibly, or moving untouched) and to ignore task success and the amount of motion.

Localization. Stage 2 follows up to 120 corner points from the first frame through the same frames with Lucas–Kanade tracking. It removes camera motion by subtracting each frame's median point velocity, then computes 16 signals from the remaining motion: the number of points, the fraction still tracked at the end, and the fraction lost away from the frame border, which is how an object vanishing mid-video shows up; the mean point speed and the 95th percentiles of speed, acceleration and jerk; seven statistics aimed at single families: the vertical acceleration of downward-moving points and its magnitude (gravity), speed after the sharpest slowdown relative to speed before it (momentum), the rate of direction reversals, the growth and fluctuation of the point cloud's spread (deformation), and centroid travel relative to rotation (friction); and the time and relative strength of the frame with the largest mean point acceleration. In the interactive tool, Stage 2 goes further. A VLM names the objects, SAM 3 segments and tracks them, Stage-1 flags are grouped into time windows, and each constraint family is ranked with a confidence and a time window. That confidence is the larger of a heuristic prior from the Stage-2 evidence and a VLM triage score. Section 4.4 uses only the 16 point-track signals.

VLM verification. Stage 3 asks each of the ten VLMs three more questions (Table 9): task completion, violation type and, on unobservable videos only, adherence to the stated property. All four questions, including the Stage-1 physics-error question, show eight frames spaced evenly from the first to the last frame, with the longer side at most 512 pixels (Llama-4-Scout receives four frames). Every question states the task. Only the hidden-property question states the property, and it also gives the scenario's expected outcome. The violation question lists eight failure descriptions and a "no physics problem" option. The eight descriptions are one for each family in Table 9 plus object permanence (an object appearing or vanishing). The options are shuffled per video with a fixed seed, so a preference for the first-listed option does not favor any family. These options are descriptions, not the human labels. There is no Contact option, so Contact is merged into Collision when scoring, and permanence has no human label.

Each answer is read from the probabilities of the VLM's first output token. We sum the probability of each allowed answer over its surface forms (e.g. "3" and " 3"), renormalize over the allowed answers, and take the expected rating for the 1–4 questions and $P(\text{yes})$ for completion. Violation detection uses $1 - P(\text{none})$, and family attribution uses $P(\text{family})$. A missing answer is replaced by that VLM's mean score. For the spread in Table 5, two further runs shift the eight frames by one third and two thirds of the spacing between them. Section 4.4 does not use the tool's specialist modules, which fit trajectories and check individual events with a VLM (Section 3.4). There, Stage 3 consists only of these questions.

Combining evidence. For each question, each VLM's scores are converted to ranks over the 439 videos, scaled to $[0, 1]$, and averaged over the ten VLMs, so that no model's calibration dominates. The 38 Stage-1–2 signals are reduced to one signal score per question. On the training folds, we keep the five signals with the largest absolute Spearman correlation with the human label (the 1–4 rating for plausibility and adherence, the binary label otherwise), flip the sign of those that correlate negatively, and average their ranks. The final score is $w\, r_{\text{VLM}} + (1 - w)\, r_{\text{sig}}$, where r denotes a rank and $w \in \{0, 0.1, \ldots, 1\}$ maximizes AUC on the training folds. The training-fold signal scores used to choose w are themselves computed out of fold, and test videos are ranked against the training videos. Signal selection and w are refit for every question and fold, using five-fold cross-validation that keeps every version of a scenario (both observability conditions and all four generators) in the same fold. Family attribution is fitted on violating videos only, and adherence on unobservable videos only.

Reporting and routing. Stage 4 writes a JSON report for each video from the shared evidence store. The report holds a chronological timeline of findings, each with its time window, object, violation type, confidence and the specialist's explanation, together with the Stage-2 family ranking and each tool's severity. An optional text LLM summarizes the report in prose. Stage 4 adds no

Table 9: VLM questions, verbatim (line breaks omitted). Each prompt follows the frames and the header; $\langle\cdot\rangle$ marks fields filled per video. The violation options appear as A)–I) in a per-video shuffled order; family names in parentheses are not shown to the model.

Question	Prompt
Header	Look at these $\langle n \rangle$ frames, sampled in order from a video of a robot. The task: "$\langle$task$\rangle$".
Physics error (plausibility)	Ignore whether the action succeeds and ignore how much motion there is. Count only PHYSICS ERRORS: objects passing through each other, appearing or vanishing, floating without support, changing shape impossibly, or moving with nothing touching them. 1 = many clear physics errors, 2 = some errors, 3 = minor oddities, 4 = no physics errors. Answer with exactly one digit, 1 to 4.
Standard plausibility (baseline)	Rate how PHYSICALLY PLAUSIBLE this video is — do objects, contact, motion, liquids and deformation behave as they would in the real world? 1 = clearly impossible physics, 2 = noticeable physics errors, 3 = minor oddities, 4 = fully plausible. Answer with exactly one digit, 1 to 4.
Task completion	Did the robot actually COMPLETE the task shown? Answer with exactly one digit: 1 = no, 2 = yes.
Hidden property (unobservable only)	Hidden property of this scene (not visible in the frames): $\langle$property$\rangle$ — $\langle$value$\rangle$. If the video respects it, this should happen: $\langle$expected outcome$\rangle$ Rate how well the video FOLLOWS that hidden property. 1 = contradicts it, 2 = mostly contradicts, 3 = mostly follows, 4 = clearly follows. Answer with exactly one digit, 1 to 4.
Violation type	(Header: Look at these $\langle n \rangle$ frames, sampled in order from a video of: "$\langle$task$\rangle$".) Which ONE of these best describes the main physics problem in this video? $\langle$options$\rangle$ Answer with exactly one letter.
Violation options	
(Collision)	Two things overlap or pass through each other instead of meeting at their surfaces, or the gripper grasps without closing on the object
(Deformation)	A rigid object changes its shape, length or thickness
(Causality)	An object moves or changes on its own, with nothing visibly touching it
(Gravity)	Something hangs in the air, floats, or fails to fall when nothing is holding it
(Momentum)	Something speeds up after being released or struck, or stops dead for no reason
(Permanence)	An object appears from nowhere or vanishes
(Fluid)	Liquid appears, vanishes, or holds an impossible rigid shape
(Friction)	Something slides when it should grip, or keeps sliding with nothing pushing it
(None)	No physics problem — everything behaves as it should

score of its own. In the interactive tool, routing happens between Stages 2 and 3: only the specialists that Stage 2 ranks highest (three by default) run on a video. Section 4.4 applies no routing, so every signal and every question is computed for all 439 videos.

Reproducibility Statement

Dataset construction, prompt schema, and the domain taxonomy are described in Appendix A. The human annotation protocol and rubric are detailed in Appendix B.

The dataset is publicly available on HuggingFace at this link: https://huggingface.co/datasets/swiftrando/PhysicsLENS.

The code for the PhysicsLENS Evaluation and Diagnostic tool is available on Github anonymously here: https://anonymous.4open.science/r/PhysicsLENS-C41F/README.md.