PHYGROUND: BENCHMARKING PHYSICAL REASONING IN GENERATIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative world models are increasingly used for video generation, where learned simulators are expected to capture the physical rules that govern real-world dynamics. However, evaluating whether generated videos actually follow these rules remains challenging. Existing physics-focused video benchmarks have made important progress, but they still face three key challenges, including the coarse evaluation frameworks that hide law-specific failures, response biases and fatigue that undermine the validity of annotation judgments, and automated evaluators that are insufficiently physics-aware or difficult to audit. To address those challenges, we introduce **PhyGround**, a criteria-grounded benchmark for evaluating physical reasoning in video generation. The benchmark contains 250 curated prompts, each augmented with an expected physical outcome, and a taxonomy of 13 physical laws across solid-body mechanics, fluid dynamics, and optics. Each law is operationalized through observable sub-questions to enable per-law diagnostics. We evaluate eight modern video generation models through a large-scale, quality-controlled human study, grounded on social science lab experiment design. A total of 459 annotators provided 5,796 complete annotations and over 37.4K fine-grained labels; after quality control, the retained annotations exhibited high split-half model-ranking correlations (Spearman's $\rho > 0.90$). To support reproducible automated evaluation, we release **PhyJudge-9B**, an open physics-specialized VLM judge. PhyJudge-9B achieves substantially lower aggregate relative bias than Gemini-3.1-Pro (3.3% vs. 16.6%). We release prompts, human annotations, model checkpoints, and evaluation code; links are available at: `https://huggingface.co/anonymouscla`.

1 INTRODUCTION

Generative world models have been increasingly applied in video generation (Li et al., 2025; Wan et al., 2025; Rupprecht et al., 2026). Ideally, world models should be capable of understanding and predicting physical phenomena (Ding et al., 2025; Li et al., 2025). However, current models often violate basic physical principles. For example, objects may float without support, fluids may disappear or lose mass, and shadows may conflict with light source. These failures undermine the physical authenticity of generated videos and limit the applicability of world models in domains requiring reliable simulation, prediction, and visual reasoning. To better understand the physical reasoning capability of current world models and identify areas for improvement, it is critical to evaluate the extent to which generated videos comply with physical laws. Such evaluation requires a reliable benchmark for measuring, comparing, and diagnosing the physical plausibility of outputs from world models. Although recent physics-focused video benchmarks have made meaningful progress in the assessment of physical adherence in generative world models, three key challenges remain unresolved.

First, existing evaluation frameworks often operate at the domain level, producing broad overall ratings that obscure law-specific failures. However, a physical domain is not a single capability. It contains multiple physical constraints that may be satisfied or violated independently. For example, a fluid scene may involve flow dynamics, boundary interaction, displacement, and continuity. A generated video may depict plausible flow while violating continuity by making liquid disappear or appear from nowhere. Averaging these distinct criteria into a single "fluid dynamics" score can therefore mask important failure modes and create a misleading impression of physical competence.

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.

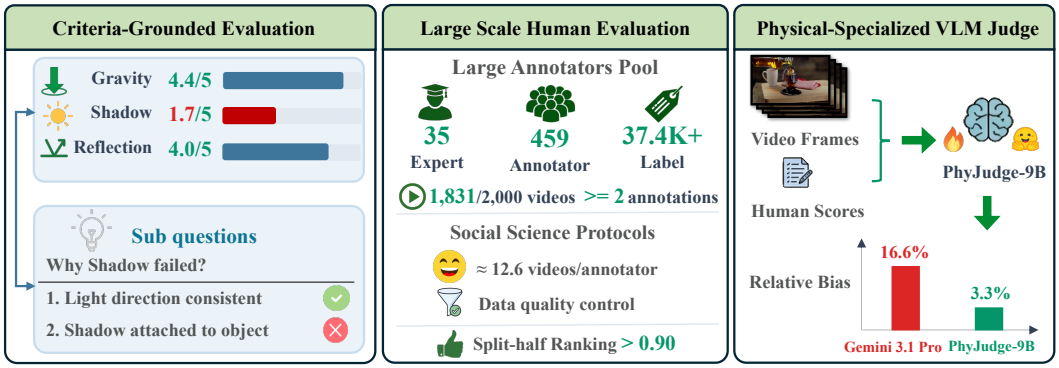


Figure 1: **Overview of PhyGround**. PhyGround decomposes each video model's holistic physical reasoning score into scores for 13 physical laws. We recruited 459 annotators to conduct a large-scale, quality-controlled human study. Based on these human annotations, we released PhyJudge-9B, a fine-tuned judge model that supports reproducible automated evaluation.

Moreover, existing frameworks often rely on binary judgments, making it impossible to distinguish differences in severity (a slightly deviated trajectory versus an object flying upward against gravity).

Second, human evaluation for physics-focused video benchmarks is vulnerable to response biases rooted in human judgment. Annotators watch videos, interpret scoring rules, identify relevant visual evidence, and assign ratings under limited time and attention. If the annotation process is not carefully designed, benchmark scores may reflect fatigue, response shortcuts, and ordering effects. These concerns are well documented in social science (Reis & Judd, 2000; Campbell & Stanley, 1963; Shadish et al., 2002; Krosnick, 1991; Orne, 1962). They are especially salient in physics-focused video evaluation, where annotators must distinguish videos that merely look plausible from those that actually satisfy the relevant physical constraints. A small annotator pool further amplifies this problem, as aggregate scores may be driven by individual biases rather than stable model performance.

Third, automatic evaluation remains insufficiently physics-aware and auditable. Standard video metrics such as FVD(Unterthiner et al., 2019), SSIM, PSNR(Hore & Ziou, 2010), and CLIP-based similarity(Radford et al., 2021) are useful for measuring distributional similarity, pixel-level fidelity, or semantic alignment, but they are not designed to detect violations of physical laws. Recent VLM-as-judge approaches provide a more semantic alternative, but many rely on closed-source models (Bansal et al., 2025; Hu et al., 2025; Li et al., 2025). Because their architectures, weights, training data, model versions, and API configurations evolve over time and are not fully controlled by benchmark users, the evaluation results may be difficult to reproduce or audit over time. Moreover, general-purpose VLM judges are not necessarily specialized for fine-grained physical diagnosis, especially when visually plausible videos contain subtle violations.

To address these challenges, we introduce **PhyGround**, a criteria-grounded benchmark for physical evaluation. As illustrated in Figure 1, it advances physics-focused video benchmarking in three ways.

First, Criteria-Grounded Per-Law Evaluation Taxonomy. We construct a criteria-grounded physics taxonomy that evaluates physical reasoning at a more fine-grained level, both for specific laws and for sub-question design. The taxonomy covers three observable domains, including solid-body mechanics, fluid dynamics, and optics, and defines thirteen per-law criteria, including laws that are often under-specified in prior benchmarks, such as shadow consistency. Furthermore, each law is operationalized through sub-questions for VLMs and scored on a 1-5 Likert scale, capturing the extent of violations. As a result, PhyGround produces per-law and per-domain diagnostic scores.

Second, Large-Scale, Quality-Controlled Human Evaluation. We conduct a large-scale human evaluation of eight video generation models using a controlled annotation protocol. A total of 459 annotators provided 5,796 complete annotations and over 37.4K fine-grained labels. Beyond scale, we design the annotation process to improve reliability. For example, drawing on social science experimental protocols, we randomize the arrival of annotators and video presentation order, cap per-annotator workloads, and apply the survey design principles to make sure to improve task clarity and question accuracy. We then filter low-quality annotation based on multiple metrics, score consistency,

dimension discriminability, peer agreement, and behavioral engagement. These procedures produce stable aggregate rankings, with split-half model-ranking correlations exceeding Spearman $\rho > 0.90$.

Third, PhyJudge-9B, A Fully Open Physics-Specialized VLM Judge. We release PhyJudge-9B, an open VLM judge fine-tuned on PhyGround human annotations. It is supervised fine-tuned on structured score supervision from human annotations and is trained for PhyGround's per-law scoring task. Compared with closed-source baseline Gemini-3.1-Pro, PhyJudge-9B achieves substantially lower aggregate relative bias (3.3% vs. 16.6%).

To summarize, PhyGround provides a fully open-source benchmark for physical evaluation. We hope to provide the community with a solid foundation for reproducible and auditable physical video evaluation. All prompts, human annotations, checkpoints, and code will be publicly released.

2 Related Work

Physics-focused Benchmarks for Video Generation. General-purpose video benchmarks such as EvalCrafter (Liu et al., 2024) and VBench (Huang et al., 2024) rely on FVD, SSIM, and CLIP-style metrics that measure visual fidelity rather than physical correctness. Recently, a growing number of physics-focused benchmarks have emerged to address this gap(Meng et al., 2024; Wang et al., 2026; Guo et al., 2025; Zhang et al., 2025a; 2026a; Wang et al., 2025c; Cao et al., 2024; Sun et al., 2025; Xu et al., 2026; Song et al., 2025). VideoPhy/-2 (Bansal et al., 2024; 2025) use a binary followed/violated rubric that collapses mild and severe violations, with about ten annotators. PhyGenBench (Meng et al., 2024) curates only 160 manually crafted prompts and assumes each video corresponds to only one physical law. Physics-IQ (Motamed et al., 2026) grades 66 scenarios with reference-based pixel metrics that presume a unique ground-truth continuation and do not localize per-law failures. WorldModelBench (Li et al., 2025) has 350 prompts with 65 annotators but its binary 0/1 scoring obscures law-specific weaknesses, easily saturates at full scores, and thus lacks the resolution to distinguish the advantages of recent video generation models. A complementary line inverts the setup: Impossible Videos (Bai et al., 2025b) prompts models to generate counterfactual scenes that violate physical, biological, or social laws, testing rule-breaking rather than rule-following. Its 260 prompts and GPT-4o-based judgments target creative generalization rather than per-law physical-correctness diagnosis. PhyWorldBench (Gu et al., 2025) and VideoScience-Bench (Hu et al., 2025) push other frontiers (anti-physics, undergraduate science) but rely on closed-source MLLM judges.

VLM-as-a-judge for Video Physics Evaluation. Standard metrics such as FVD(Unterthiner et al., 2019), SSIM(Hore & Ziou, 2010), and CLIP(Radford et al., 2021) measure distributional similarity or semantic alignment, not violations of physical laws. Learned video-quality metrics like VideoScore (He et al., 2024) score fine-grained quality dimensions but are not trained on physics-specific labels. Recent benchmarks pair their datasets with VLM-based automatic evaluators (Zheng et al., 2025; Motamed et al., 2025; Huang et al.; Han et al., 2025; Liu & Zhang, 2025; Wang et al., 2025a; He et al., 2025; Liu et al., 2025; Tong et al., 2025; Wang et al., 2025b; Zhang et al., 2025b; 2026b). On the open side, VideoPhy (Bansal et al., 2024) fine-tunes VideoCon-Physics, VideoPhy-2 (Bansal et al., 2025) releases VideoPhy-2-AutoEval, and WorldModelBench (Li et al., 2025) releases a 2B judger reported to outperform GPT-4o by 8.6% on world-modeling violations. These open evaluators are trained on holistic single-score labels and architecturally cannot produce per-law diagnostics or severity calibration even when reproducible. A second line uses closed-source VLMs: PhyGenEval (Meng et al., 2024), VideoScience-Bench (Hu et al., 2025), and PhyWorldBench (Gu et al., 2025) all rely on GPT-4o or Gemini variants, raising reproducibility concerns because model versions, weights, and API behavior change outside benchmark users' control.

Positioning of PhyGround. Table 1 compares PhyGround with six physics-focused video generation benchmarks. First, PhyGround supports more diagnostic evaluation by combining explicit expected-outcome prompts with per-law decomposition. Second, PhyGround provides a more reliable human-evaluation foundation through a substantially larger quality-controlled annotator pool. Third, PhyGround enables reproducible automated evaluation by providing an open-source judge.

Table 1: Comparison of physics-focused video generation benchmarks. **Laws**: number of distinct physics laws checked at evaluation (– if not score at law granularity). **Annotators**: number of human annotators reported for evaluation (– if not reported). **Video/Ann.** ↓: average number of generated videos labeled per annotator (workload per annotator; lower means each annotator focuses on fewer videos). **Explicit**: prompts state the expected physical outcome. **Sub-Q**: automatic judge prompt includes law-specific sub-questions. **Open Judge**: open-source evaluator is provided.

Benchmark	Prompts	Videos	Laws	Annotators	Video/Ann. ↓	Explicit	Sub-Q	Open Judge
PhyGenBench	160	1,280	27	3	512	✗	✓	✗
Physics-IQ	66	396	–	–	–	✗	✗	✓
VideoPhy-2	688	6,800	19	12	1,700	✗	✓	✓
VideoScience-Bench	200	–	–	–	–	✓	✓	✗
WorldModelBench	350	4,421	5	65	128	✗	✗	✓
Impossible Videos	260	902	6	–	–	✗	✗	✗
PhyGround (Ours)	**250**	**2,000**	**13**	**459**	**12.6**	✓	✓	✓

3 Benchmark Design

3.1 Prompt Suite Construction

For objective, explicit evaluation of physical correctness, we curate 250 prompts from four complementary sources and retain those with unambiguous physical outcomes, visually verifiable events, and concise everyday scenarios (criteria and rationale in Appendix A.1).

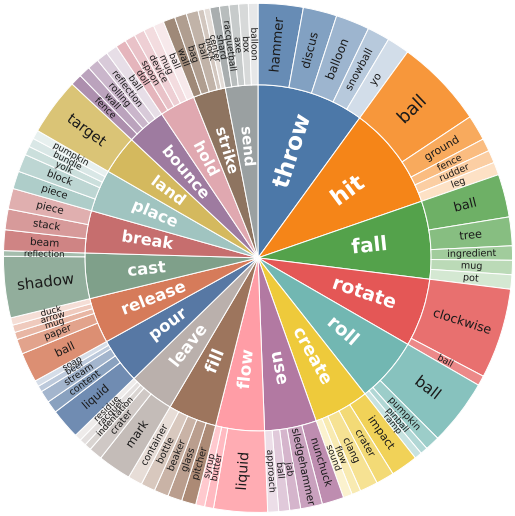


Figure 2: Distribution of root verbs (inner ring, top 20 by frequency) and their direct nouns (outer ring, top 5 per verb) across the 250 PhyGround prompts.

Expected-outcome Augmentation. Previous benchmarks were mostly based on implicit prompts: they described only the process, without specifying the physical outcome, such as a ball falling or a ball bouncing. The original intention was to see whether video models could infer the outcome. However, many physical phenomena do not necessarily occur, which allows the video to have multiple possible endings and greatly affects the reproducibility of the evaluation. Implicit prompts can tell whether a generated video looks plausible vaguely, but they do not always reveal whether the expected physical outcome actually occurs. Therefore, we revise each implicit prompt by adding an explicit expected physical outcome. For instance, "A ball rolls across a table" becomes "A ball rolls across a table and *reaches the edge, falls off*." This forces the model to attempt the physical phenomenon, raising the bar on physical correctness and giving evaluators a clear scoring reference. PhyGround therefore measures outcome-conditioned physical adherence: whether a model realizes a specified physical event plausibly, judged as structured perceived adherence to law-specific observables rather than by instrumented measurement of trajectories or conserved quantities (Appendix L). All prompts use a text+image-to-video (TI2V) format with the first frame as image condition.

Curating Procedure. First frames are drawn from existing video corpora, in particular, WorldModelBench and Physics-IQ provide reference frames directly, OpenVid and VideoPhy-2 first frames are extracted as the first frame of their source video. Because some source videos can contain content that conflicts with our grading criteria, we manually verify that every first-frame image and text condition is of good quality before we include it in the prompt suite.

Prompt Diversity. Figure 2 illustrates the distribution of root verbs and direct nouns across the 250 prompts. The 250 prompts cover a broad and well-balanced range of physical actions and entities: no single verb dominates, and each verb fans out into multiple distinct objects spanning rigid-body motion (throw/hit/roll/bounce), gravity and contact (fall/land/place/break), fluids (pour/flow/fill), optics (cast/shadow/reflection), and tool use (use, hold, strike).

3.2 Physics Criteria Taxonomy

We define a curated set of thirteen physical laws organized across three domains. In particular, we decompose "fluid dynamics" into flow dynamics, boundary interaction, displacement, and continuity, allowing different physical constraints within the same domain to be diagnosed independently. Each video is evaluated with 2 to 4 applicable laws based on its prompt content. The selection criteria for laws and the excluded categories are detailed in Appendix A.2.

Following social science research design principles on validity and measurement quality (Reis & Judd, 2000), we use a 1-5 scale to preserve differences in violation severity, where 1 indicates that fully implausible and 5 indicates fully plausible. This approach avoids the limitation of a binary measure, which would collapse all deviations into the same category and obscure the distinction between minor imperfections and severe physical failures.

3.3 Automatic VLM Judge

PhyGround provides an automatic VLM judge for reproducible large-scale evaluation. The judge takes the video frames, the augmented prompt, and a criterion-specific rubric as input, and outputs a 1-5 score for one general dimension or one applicable physical law. For physical-law scoring, the same per-law sub-questions are included in the judge prompt as an observation checklist. In practice, we query the judge separately for each criterion, each video requires 5–7 VLM queries in total: three for the general dimensions and 2-4 for the applicable physical laws. This criterion-wise querying keeps each prompt focused on a single rubric, reduces interference among heterogeneous criteria, and produces more reliable, easily parsed diagnostic scores.

Sub-question Decomposition. For the VLM judge, we equip each law with 2-3 sub-questions (SubQ), each targeting a specific observable property. These sub-questions operationalize an abstract law into visually verifiable observation items and are explicitly listed in the VLM judge's prompt as a reasoning checklist that constrains its observation perspective and reduces subjectivity.

3.4 Scoring

General Physical Adherence We evaluate each video along three complementary dimensions. **Semantic Alignment (SA)** measures the extent to which the video content matches the text prompt. **Physical Temporal Validity (PTV)** assesses whether the temporal sequence of physical events follows a plausible physical order. **Object Persistence (Persistence)** evaluates whether objects maintain consistent appearance, shape, and existence throughout the video.

Per-law Physics Scores. For each video, we evaluate only the subset of physical laws applicable to its prompt, resulting in 2-4 law-level scores per video. Each law is scored independently, with the score reflecting the degree to which the generated video follows that law.

Human Annotation. We design the annotation process using established principles for valid and reliable human measurement (Campbell & Stanley, 1963; Reis & Judd, 2000; Shadish et al., 2002). Annotators complete the task on desktop devices only to ensure consistent viewing conditions and reduce variation in screen size, resolution, and interface layout. To limit fatigue and low-effort responding (Krosnick, 1991), each annotator rates roughly 12.6 videos on average. The task is described without revealing our hypotheses, and model identities are hidden to reduce demand effects (Orne, 1962). After consenting, participants report basic demographic information and complete a short training module that illustrates different levels of physical realism and explains the five-point Likert scale. During evaluation on our annotation platform, each participant is randomly assigned videos from the full pool, and video order is randomized independently to reduce assignment bias,

ordering effects, and carryover effects. Each video must be watched at least once before rating. Annotators score three general dimensions and 2-4 physical laws, each on a 1-5 Likert scale. Finally, we apply a two-round quality-control pipeline based on score consistency, dimension discriminability, peer agreement, and behavioral engagement.

3.5 JUDGE TRAINING

We perform supervised fine-tuning with LoRA adapters, with training data drawn from the human annotation platform (Section 4). Each sample is a triplet (v, q_k, y_k): video frames v, the augmented prompt q_k instantiated for a single criterion k (containing the expected outcome), and the structured JSON score label $y_k = \{k : s\}$. Here, $s \in \{1, \ldots, 5\}$ denotes the score obtained by aggregating human annotations. Let the dataset be $\mathcal{D}$; only the LoRA parameters θ are trainable, and the training objective minimizes the token-level negative log-likelihood:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(v,q_k,y_k)\sim\mathcal{D}} \left[\sum_{t=1}^{|y_k|} \log p_\theta\big(y_{k,t} \mid y_{k,<t}, v, q_k\big) \right]. \tag{1}$$

Here $y_{k,t}$ denotes the t-th token of the serialized JSON label y_k, and $|y_k|$ is its token length. Training uses a holdout validation strategy, we evaluate the model on Veo-3.1 hold-out split and train on the remaining 7 models human annotations. The LoRA configuration, score-supervision format, and training hyperparameters in Appendix H. We refer to the resulting fine-tuned judge as **PhyJudge-9B**.

4 EXPERIMENTS

4.1 SETUP

Video Generation Models. We evaluate 8 video generation models spanning a representative range of architectures and scales, including one closed-source model (Veo-3.1 (Google DeepMind, 2026)) and seven open-source models (Wan2.2-27B-A14B (Wan et al., 2025), OmniWeaving (Pan et al., 2026), Cosmos-Predict2.5-14B and Cosmos-Predict2.5-2B (Gu, 2025), Wan2.2-TI2V-5B (Wan et al., 2025), LTX-2.3-22B and LTX-2-19B (HaCohen et al., 2026)). All models are evaluated under the unified ti2v format with 250 prompts, producing 2,000 total videos. Unless otherwise stated, each model is evaluated under its native default generation settings, including output resolution and frame rate, without post-hoc normalization (see Appendix K for details).

Evaluation Design. In total, we recruited 459 annotators, who in total contributed 5,796 complete annotations and 37.4K individual labels through our evaluation platform, with all 2,000 videos. Each video receives at least two annotators (1,410 with at least three). To the best of our knowledge, this is the largest human-annotator pool reported for a physics-focused video generation benchmark (Table 1). For each law, the rubric anchor shown to annotators is the per-law main question together with its accompanying note (e.g., for gravity: "unsupported objects fall downward, projectiles follow curved trajectories, poured liquids descend"), keeping the annotation interface concise. After two rounds of quality control, cross-validating four mutually independent signals (score consistency, per-video dimension distinctness measured by copy-paste rate, peer MAE, and behavioral signals. Full procedure in Appendix B) 352 annotators are retained, yielding 4,576 complete annotations and 29.5K individual labels, with 1,831 of the 2,000 videos still covered by at least two annotators.

For VLM evaluation, we include Claude (Opus 4.7), Gemini-3.1-Pro(Team et al., 2023), Qwen3.5 9B-base, and Qwen3.5 27B-base (Bai et al., 2025a), with both direct-answer and Chain-of-Thought prompting variants where applicable, as baseline models in our setup.

4.2 MAIN RESULTS

Table 2 shows the scores for all models. Each general dimension (SA, PTV, Persist.) is averaged per video across annotators and then across videos. Each physics-domain score (Solid-Body, Fluid, Optical) is the mean over all (video, law) annotation units in that domain, so each law is weighted by the number of videos it applies to. Overall averages the General and Physics scores: $0.5 \times$ mean(SA, PTV, Persist.) $+ 0.5\times$ the (video, law) mean pooled across the three physics domains. At

Table 2: Human evaluation results on PhyGround (1–5 scale, higher is better) over 250 prompts. SA, PTV, and Persist. denote Semantic Alignment, Physical Temporal Validity, and Persistence; Solid-Body, Fluid, and Optical denote the three physics-domain scores. Overall combines the General and Physics views with equal weight. **Bold** marks the best score in each column and underline marks the second-best. †: closed-source.

	General Physics ↑			Domain Physics ↑			
Model	SA	PTV	Persist.	Solid-Body	Fluid	Optical	Overall ↑
Wan2.2-27B-A14B	<u>3.10</u>	**3.37**	**3.50**	**3.23**	3.18	<u>3.55</u>	**3.28**
Veo-3.1†	**3.26**	<u>3.29</u>	<u>3.42</u>	<u>3.12</u>	**3.65**	**3.69**	**3.28**
OmniWeaving	2.97	3.17	3.34	2.98	<u>3.26</u>	3.22	3.10
Cosmos-14B	2.66	2.98	3.20	2.81	2.98	3.36	2.90
LTX-2.3-22B	2.58	2.74	2.72	2.62	3.07	2.98	2.69
Wan2.2-TI2V-5B	2.44	2.68	2.78	2.58	2.71	2.99	2.63
Cosmos-2B	2.33	2.56	2.77	2.53	2.77	3.35	2.58
LTX-2-19B	2.56	2.64	2.48	2.46	3.04	2.78	2.56

a high level, no model achieves strong physics adherence: no model exceeds 3.3/5 (66%) on Overall, and no model crosses this threshold on the solid-body domain either, indicating substantial room for improvement in physical reasoning. However, reading only Overall provides an incomplete picture, motivating the cross-model analysis along the Solid-Body / Fluid / Optical columns presented below.

Per-domain decomposition reveals complementary model profiles. Holistic rankings can obscure substantial domain-level differences. For example, the top open-source Wan2.2-27B-A14B and the closed-source Veo-3.1 obtain the same Overall score (both 3.28), yet their strengths differ sharply: Wan2.2-27B-A14B matches or exceeds Veo-3.1 on 4 of the 7 evaluation dimensions, achieving the best scores in the table on solid-body (3.23 vs. 3.12), temporal coherence (PTV, 3.37), and persistence (3.50), whereas Veo-3.1 leads on fluid (3.65 vs. 3.18, a gap of 0.47, bootstrap 95% CI $[0.27, 0.68]$) and optical (3.69 vs. 3.55). A similar effect appears near the bottom: Cosmos-Predict2.5-2B and LTX-2-19B differ by only 0.02 on Overall (2.58 vs. 2.56), but the former is weak on fluid (2.77) and strong on optical (3.35), while the latter is competitive on fluid (3.04), has the lowest solid-body score in the table (2.46), and is weak on optical (2.78). Thus, models with nearly identical holistic scores may imply very different choices for deployment, model selection, and targeted improvement.

This decomposition also exposes cases where a low holistic score hides a functioning sub-capability. Although Cosmos-Predict2.5-2B ranks only 7th on Overall, its optical score reaches 3.35, essentially tied with Cosmos-Predict2.5-14B (3.36) and exceeding LTX-2.3-22B, Wan2.2-TI2V-5B, and LTX-2-19B (all < 3.0). Its within-model range, from solid-body 2.53 to optical 3.35, is the largest among the 8 evaluated models and crosses nearly a full rating level, showing that its Overall score is pulled down by uneven domain performance rather than uniformly weak behavior across all domains. Per-law sample counts, bootstrap CIs and split-half stability for every law are reported in Appendix G; the two Cosmos-2B / LTX-2-19B contrasts above are significant at the 95% level (Table 23).

Finally, aggregating scores by physics domain rather than by model reveals a benchmark-level difficulty profile: the cross-model means are 2.79 on solid-body, 3.08 on fluid, and 3.24 on optical. This suggests that scenarios involving rigid-body motion, force transmission, and collision response remain particularly challenging.

The value of the per-domain decomposition is not that it correlates better with human ratings than holistic scoring, but that **it extracts an information dimension that holistic scoring is structurally unable to contain**. Appendix D further drills down to per-law granularity, revealing finer-grained observations that are averaged out even by the three-domain aggregation.

4.3 Annotation Reliability

Leaderboard Stability. We assess ranking robustness by measuring score shifts under annotator subsampling. Dropping a random 30% of annotators changes per-model scores by only 0.033 RMS, while sub-sampling down to 20 annotators inflates this to 0.204 RMS, a sixfold gap that highlights

Table 3: Split-half ranking reliability of the leaderboard. The 352-annotator post-QC pool is randomly partitioned into two equal halves; each cell reports mean ± standard deviation over 100 random partitions.

Score	Spearman's ρ	Kendall τ-b
General	0.930 ± 0.045	0.828 ± 0.071
Physics	0.907 ± 0.058	0.801 ± 0.092
Overall	0.928 ± 0.051	0.830 ± 0.088

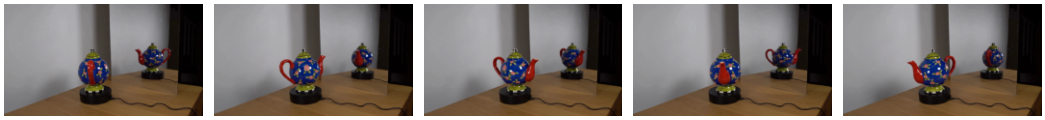
Table 4: Pairwise preference agreement on General, Physical, and Total scores, with WorldModelBench and VideoPhy-2 shown as reference points. Ties are excluded; N is the number of duplicate-vote comparisons after excluding ties.

Dimension	Pairwise agreement (%) ↑	N
General (avg)	76.0	51,872
Physical (avg)	74.5	51,617
Total score	73.0	58,984
WorldModelBench	70.0	—
VideoPhy-2	75–80	—

the advantage of using a larger candidate pool (full sweep in Appendix B). In a split-half check, we partition the 352 annotators into two equal groups and compare the resulting model rankings over 100 random partitions. The halves agree strongly on General/Physics/Overall (Table 3). Thus, while individual annotators may disagree on specific videos, the aggregate benchmark reliably distinguishes model quality. The ranking is likewise unchanged ($\rho \geq 0.98$) when the reference is restricted to education, major or cohort subgroups of the pool (Appendix J.7).

Inter-Annotator Agreement. We compute same-prompt pairwise preference agreement using the same win/loss-style metric as WorldModelBench (Li et al., 2025) to summarize inter-annotator reliability. Physical (avg) averages each annotator's per-law scores into a single physics score per video before comparing, and Total score is General (avg) + Physical (avg). Table 4 reports this agreement on the General/Physical/Total scores. Agreement is similar for the General and Physical scores (76.0% vs. 74.5%), while the Total score obtains 73.0%. These values are comparable to those reported by WorldModelBench (70%) and VideoPhy-2 (75-80%).

4.4 Judge Results



Domain: Optical Model: Veo3.1

Prompt: A teapot on a rotating display base in front of a mirror; the reflection rotates in the opposite direction.

Judge	SA	PTV	Persist.	Reflection	Shadow	Overall
Human	4.0	4.3	4.3	4.7	4.3	4.4
PhyJudge	4	5	5	4	5	4.6
Gemini	3	5	2	1	1	2.2

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Reflection (Gemini=1, Human=4.7):
"when the real spout points towards the mirror, the reflected spout points away from it, violating the basic laws of reflection. The physical behavior is severely incorrect"
Shadow (Gemini=1, Human=4.3):
"The shadow of the teapot is baked into the object/base and rotates with it, completely violating the stationary light source"

Figure 3: Head-to-head comparison of PhyJudge-9B and Gemini-3.1-Pro on a representative case. Green cells mark dimensions where PhyJudge-9B tracks the human mean more closely than Gemini; red cells mark dimensions where Gemini diverges from the human mean by at least two rating levels. The panel beneath reproduces Gemini's verbatim conclusion sentence on each disputed law.

The main judge results are reported in Table 5 (judge per-dimension aggregate relative bias). PhyJudge-9B's aggregate relative bias is lower than that of the closed-source Gemini-3.1-Pro and Claude (Opus 4.7). The controlled comparison is the same backbone without fine-tuning: Qwen9B-base scored with the identical prompt sits at 31.6%, so fine-tuning is what buys the calibration,

Table 5: Per-dimension aggregate relative bias (%) with per-law sub-question prompts; CoT rows use reasoning-augmented outputs. The first three columns report $E = |\bar{J} - \bar{H}|/\bar{H}$ per dimension/law, macro-averaged; Physics is macro-averaged over domains and Overall = 0.5·General + 0.5·Physics. The last column is the WorldModelBench (Li et al., 2025) signed bias $(\bar{J} - \bar{H})/\bar{H}$ on the per-video overall score (closer to zero is better). **Bold** marks the best value per column.

Model	**General (%)** ↓	**Physics (%)** ↓	**Overall (%)** ↓	**Signed (%)** → 0
Claude (Opus 4.7)	16.2	10.0	13.1	+14.5
Claude (Opus 4.7, CoT)	25.2	34.4	29.8	+34.3
Gemini-3.1-Pro	17.3	15.9	16.6	-6.3
Gemini-3.1-Pro (CoT)	17.2	17.4	17.3	-6.5
Qwen9B-base	28.2	35.0	31.6	+37.7
Qwen9B-base (CoT)	41.8	41.6	41.7	+47.8
Qwen27B-base	21.3	40.9	31.1	+35.7
Qwen27B-base (CoT)	37.3	46.9	42.1	+48.3
PhyJudge-9B	**1.0**	**5.5**	**3.3**	**-1.1**

Table 6: Raw Gemini-3.1-Pro judge means under prompt-schema ablations on the human-eval set. Each cell reports $\bar{J}$ with absolute relative bias $|\bar{J} - \bar{H}|/\bar{H}$ in parentheses, pooled over law-video pairs in the domain. Adding SubQ closes large gaps to human means on Fluid and Optical; CoT on top of SubQ is not consistently beneficial. **Bold** marks the schema closest to $\bar{H}$ in each row.

Domain	$\bar{H}$	**Gemini-3.1-Pro** ↓	**+ CoT** ↓	**+ SubQ** ↓	**+ SubQ + CoT** ↓
Solid-Body	2.95	1.76 (40%)	2.02 (31%)	2.29 (22%)	**2.65 (10%)**
Fluid	3.57	1.86 (48%)	1.99 (44%)	**3.29 (8%)**	2.84 (21%)
Optical	3.62	2.64 (27%)	2.68 (26%)	3.28 (9%)	**3.68 (2%)**

and it improves every per-video metric as well (MAE 1.29 → 0.79, within-±1 44.4% → 68.8%; Appendix J.1). Aggregate bias measures calibration only; Appendix J reports per-video error for all judges, ranking at the (generator, law) granularity, eight leave-one-generator-out folds, transfer to VideoPhy-2, leave-one-law-out folds, and sensitivity to the human reference. Our evaluation protocol is strictly held-out: all PhyJudge-9B tests are conducted on Veo-3.1 videos, and the Veo-3.1 videos and their human ratings do not appear in the training set of PhyJudge-9B. All judges are evaluated with the same set of per-law sub-questions. Rows marked "CoT" use a reasoning-augmented prompt: the judge first answers each sub-question with brief evidence and a written justification before emitting the final integer score, whereas the non-CoT rows constrain the same model to a score-only output. Table 6 gives the raw-score view behind this design choice: adding sub-questions moves Gemini-3.1-Pro much closer to human means on the motion-heavy Fluid and Optical domains ($1.86 \rightarrow 3.29$ and $2.64 \rightarrow 3.28$), while adding CoT on top of SubQ is not consistently beneficial.

A representative head-to-head case. Figure 3 drills into a single Optical case—a teapot on a rotating display in front of a mirror, generated by Veo-3.1. Humans rate the clip 4.4 because reflection and shadow track the teapot correctly, and PhyJudge-9B closely follows with 4.6, while Gemini-3.1-Pro scores both reflection and shadow at 1 and misreads a physically plausible scene as a near-total optical violation. Both use the sub-question prompt (Gemini-3.1-Pro with CoT, PhyJudge-9B score-only); examples for all other laws are in Appendix E. All main results use 4 fps; sweeping 1/2/4/8 fps changes scores little and without directional bias (Table 26).

5 Conclusion

We introduce PhyGround, a criteria-grounded benchmark for evaluating physical reasoning in generative world models at the level of individual physical laws. It reveals that video generation models can exhibit distinct strengths and weaknesses across different physical constraints. We further provide a large-scale, quality-controlled human annotation set and an open physics-specialized automatic judge to support scalable and reproducible evaluation. Overall, this work advances our understanding of current video generation models as world models and provides a diagnostic foundation for future improvements in physical reasoning.

AI USE STATEMENT

Generative AI is used in this work in three distinct roles, all of them disclosed in the paper. First, generative video models are the object of study: the systems compared in Section 4 produce the videos that PhyGround scores. Second, generative AI is part of the benchmark pipeline. Gemini 2.5 Flash augments each implicit prompt with a short description of the expected physical outcome (Appendix A.2), and the automatic scores come from VLM judges: PhyJudge-9B, which we fine-tune and release (Section 3.5), and the Gemini-3.1-Pro, Claude (Opus 4.7), and Qwen3.5 baselines (Section 4.1). Every augmented prompt was manually checked against the first frame and the filtering criteria before entering the prompt suite, and every judge score we report is measured against the human ratings collected in our annotation study. Third, we used LLM-based assistants for language editing and for LaTeX formatting of this manuscript. The research questions, the benchmark design, the experiments, and the analyses are the authors' own. We have reviewed all AI-assisted work, and we take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

- **Benchmark.** The prompt suite, the 13-law taxonomy and the scoring dimensions are specified in Section 3.1, Section 3.2 and Section 3.4. Appendix A gives how prompts are sourced, filtered and augmented with the expected outcome, and the per-law sub-questions; the licenses of all four prompt sources, which permit research use and redistribution, are in Appendix A.1.2.
- **Video generation.** The inference settings of every video model (resolution, FPS, frames, CFG scale, sampling steps, precision) are listed in Table 37, and every open model is run at its released defaults (Appendix K).
- **Human study and statistics.** Recruitment, the annotation interface, and the two-round quality-control criteria for removing annotators are described in Appendix B. The leaderboard aggregation and its sensitivity to the weighting scheme are in Appendix C; the bootstrap and split-half procedures behind every confidence interval are in Section 4.3, Appendix B and Appendix G.
- **PhyJudge-9B.** The judge is specified in Section 3.5; its LoRA, optimizer and compute configuration are in Table 24, its training-data format in Appendix H, and the judge prompts, frame sampling and baseline-judge settings in Appendix I. The released judge has fixed weights and runs on a single GPU without any closed-source API call, so its scores do not drift with provider-side model updates.
- **Released artifacts.**
  - Data (benchmark prompts with their first frames, all 2,000 generated videos, human annotations): `https://huggingface.co/datasets/anonymouscla/phyground`.
  - Code (data processing, judge training, evaluation and scoring, with the training configurations): `https://huggingface.co/datasets/anonymouscla/phyground-code`.
  - PhyJudge-9B LoRA adapter, with its training arguments and an inference script: `https://huggingface.co/anonymouscla/phyjudge-9B`.

REFERENCES

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*, 2025a.

Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. *arXiv preprint arXiv:2503.14378*, 2025b.

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. *arXiv preprint arXiv:2406.03520*, 2024.

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.

Donald T. Campbell and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago, IL, 1963.

Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. Physgame: Uncovering physical commonsense violations in gameplay videos. *arXiv preprint arXiv:2412.01800*, 2024.

Jesse Chandler, Pam Mueller, and Gabriele Paolacci. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46 (1):112–130, 2014.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Google DeepMind. Veo 3. `https://deepmind.google/models/veo/`, 2026.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, et al. " phyworldbench": A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*, 2025.

Jinwei Gu. Cosmos world foundation models for physical ai. In *Proceedings of the 3rd International Workshop on Rich Media With Generative AI*, pp. 39–39, 2025.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Lanxiang Hu, Abhilash Shankarampeta, Yixin Huang, Zilin Dai, Haoyang Yu, Yujie Zhao, Haoqiang Kang, Daniel Zhao, Tajana Rosing, and Hao Zhang. Benchmarking scientific understanding and reasoning for video generation using videoscience-bench. *arXiv preprint arXiv:2512.02942*, 2025.

Yujie Huang, Wenwu He, Congcong Liu, Dong Liang, and Zhuo-Xu Cui. Cosmos-eval: Towards explainable evaluation of physics and semantics in text-to-video models.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Jon A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991. doi: 10.1002/acp.2350050305.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Xiao Liu and Jiawei Zhang. Aigve-macs: Unified multi-aspect commenting and scoring model for ai-generated video evaluation. *arXiv preprint arXiv:2507.01255*, 2025.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 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. *arXiv preprint arXiv:2410.05363*, 2024.

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility. *arXiv preprint arXiv:2510.07550*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026.

Martin T. Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11):776–783, 1962. doi: 10.1037/h0043424.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, et al. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Harry T. Reis and Charles M. Judd (eds.). *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, Cambridge, UK, 2000.

Timothy Rupprecht, Pu Zhao, Amir Taherin, Arash Akbari, Arman Akbari, Yumei He, Sean Duffy, Juyi Lin, Yixiao Chen, Rahul Chowdhury, et al. Human cognition in machines: A unified perspective of world models. *arXiv preprint arXiv:2604.16592*, 2026.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, MA, 2002.

Tingyu Song, Tongyan Hu, Guo Gan, and Yilun Zhao. Vf-eval: Evaluating multimodal llms for generating feedback on aigc videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21126–21146, 2025.

Shangkun Sun, Xiaoyu Liang, Bowen Qu, and Wei Gao. Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency. *arXiv preprint arXiv:2502.04076*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, Shi Qiu, Siwei Han, Kexin Geng, Zhongkai Xue, Yiyang Zhou, Peng Xia, et al. Mj-video: Fine-grained benchmarking and rewarding video preferences in video generation. *arXiv preprint arXiv:2502.01719*, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

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.

Jiarui Wang, Huiyu Duan, Ziheng Jia, Yu Zhao, Woo Yi Yang, Zicheng Zhang, Zijian Chen, Juntong Wang, Yuke Xing, Guangtao Zhai, et al. Love: Benchmarking and evaluating text-to-video generation and video-to-text interpretation. *arXiv preprint arXiv:2505.12098*, 2025a.

Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18869–18880, 2025b.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, et al. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*, 2026.

Zeqing Wang, Keze Wang, and Lei Zhang. Phydetex: Detecting and explaining the physical plausibility of t2v models. *arXiv preprint arXiv:2512.01843*, 2025c.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025a.

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. *arXiv preprint arXiv:2603.19607*, 2026a.

Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. Vq-insight: Teaching vlms for ai-generated video quality understanding via progressive visual reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12870–12878, 2026b.

Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. Q-bench-video: Benchmark the video quality understanding of lmms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3229–3239, 2025b.

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 Prompt Suite and Physics Law Selection Rationale

A.1 Prompt Selection

A.1.1 Suite Composition

Table 7 gives the number of prompts retained from each source after the filtering below.

Table 7: Composition of the prompt suite by source.

Source	Prompts
VideoPhy-2	166
Physics-IQ	37
OpenVid	35
WorldModelBench	12
Total	**250**

A.1.2 Source Licenses

All four prompt sources are released under permissive licenses that allow research use and redistribution. VideoPhy-2 is released under the MIT License, while WorldModelBench, Physics-IQ, and OpenVid are released under the Creative Commons Attribution 4.0 (CC BY 4.0) license. We redistribute only the prompts and first-frame images from these sources, with attribution.

A.1.3 Filtering Criteria

A prompt is retained only if it simultaneously satisfies all of the following:

1. **Unique Physical Outcome.** The expected outcome must be unambiguous, admitting only one physically plausible result. For example, "spaghetti breaks into pieces" is excluded because the number of pieces has no correct answer.
2. **Visual Verifiability.** Whether the outcome occurred must be immediately apparent from the generated video, without domain expertise or frame-by-frame analysis.
3. **Concise Everyday Scenario.** The prompt must describe a concrete situation in under 50 words, set in a familiar context. Prompts requiring specialized laboratory equipment or rare objects (e.g., "rotary filling machine," "press brake") are excluded because video models lack sufficient visual reference for these objects.
4. **Domain Alignment.** The core phenomenon must belong to the target physics domain. Prompts only superficially associated through keywords are excluded. For example, grinding should not be classified as collision when the mechanism is friction, nor a penguin entering water as fluid dynamics when the scene mainly shows animal behavior.
5. **Physics as Core Phenomenon.** The scenario must be driven by a natural physical process. Machine and tool operations (hydraulic lifts, excavators, mechanical presses) are excluded because the physical outcome is incidental to the mechanical action. Active human or animal motion (jumping, throwing, running) is excluded because the initial conditions (applied force, launch angle) are unobservable, making the space of correct outcomes unbounded. For example, "The athlete throws a javelin" does not specify whether it should travel 30 or

60 meters. This contrasts with passive processes (free fall, fluid flow, collision rebound), where the outcome is fully determined by the observable initial state.

A.1.4 EXPECTED-OUTCOME AUGMENTATION

Many original prompts describe only the initial setup without stating the expected physical result. For instance, "The grabber tools let go of the ball and block" does not mention that the objects should fall. Such prompts let the model avoid the physical event entirely, making evaluation vacuous.

We use Gemini 2.5 Flash to augment each implicit prompt by appending a concise description of the expected outcome (typically 10–20 words), without altering the scene semantics; each augmented prompt is then manually checked for consistency with the first frame and filtering criteria. Examples:

- "A ball rolls across a table" → "*... reaches the edge, falls off, and bounces on the floor below.*"
- "Water is poured into a cup on a baking tray" → "*... until the cup overflows, with water spilling onto the baking tray surface.*"
- "The wooden stick hits the first block as it rotates" → "*... The blocks topple over one by one in a domino chain reaction.*"

For OpenVid, whose source data consists of lengthy video captions (often over 100 words), captions are rewritten into concise physics-focused prompts (under 50 words) with the physical event made explicit, grounded against the source video's first frame for visual consistency.

Prompts that already specify the physical outcome are left unchanged.

A.2 PHYSICS LAW SELECTION

We build the 13-law taxonomy from the physics indicators of PhyGenBench (Meng et al., 2024), Physics-IQ (Motamed et al., 2026), VideoPhy-2 (Bansal et al., 2025), VideoScience-Bench (Hu et al., 2025), and WorldModelBench (Li et al., 2025). These benchmarks organize physics along different axes: PhyGenBench uses four broad domains, Physics-IQ and VideoPhy group by scenario type, VideoScience-Bench uses a three-level university-course hierarchy, and WorldModelBench lists specific physical laws. Three selection criteria consolidate them into one vocabulary.

A.2.1 SELECTION CRITERIA

Visual Observability. The phenomenon must be immediately verifiable from a short generated video. This excludes quantities not directly visible (temperature, energy, electromagnetic fields) and phenomena like melting and boiling that rarely unfold within a short generated video.

VLM Discriminability. Each law must map to concrete visual questions a VLM judge can reliably answer (e.g., "Do unsupported objects fall downward?" not "Does the video obey gravity?").

Paradigm Compatibility. The evaluation must work without a ground-truth reference video (the TI2V setting of Section 3.1). This excludes Physics-IQ's pixel-level metrics (Spatial IoU, Spatiotemporal IoU, Weighted Spatial IoU, MSE), which assume a unique ground-truth continuation and penalize physically plausible but visually different generations.

A.2.2 INCLUDED LAWS

Table 9 lists the three general criteria and the thirteen laws with the question each puts to annotators. Table 8 lists the number of (prompt, law) pairs per law across the 250-prompt suite, 876 pairs in total; this is the population behind every per-law and per-domain mean reported in the paper.

Solid-body Mechanics (6 laws). Gravity, inertia, momentum conservation, impenetrability, collision response, and material properties. These represent the most fundamental everyday physical commonsense and produce the clearest visual evaluation criteria. Solid-body mechanics laws form the main body of the benchmark because nearly every video generation scenario involves rigid objects.

Table 8: (Prompt, law) pairs per law in the prompt suite, scored for every leaderboard model.

Domain	Law	Pairs
Solid-Body	Gravity	123
	Inertia	55
	Momentum	123
	Impenetrability	117
	Collision	162
	Material	113
Fluid	Buoyancy	18
	Displacement	26
	Flow dynamics	32
	Boundary interaction	32
	Continuity	28
Optical	Reflection	17
	Shadow	30

Fluid (5 laws). Buoyancy (dense objects sink, lightweight objects float), displacement (liquid level rises upon submersion, containers overflow when full), flow dynamics (bulk liquid motion patterns), boundary interaction (splashing and deflection at obstacles), and fluid continuity (mass conservation, no spontaneous gaps or generation). We merge buoyancy and displacement into the fluid domain because both are governed by liquid behavior and share the physical substrate of flow dynamics; treating them as a separate domain left too few (video, law) units for a stable domain-level mean.

Optical Physics (2 laws). Reflection (mirrored content matches the scene) and shadow (direction consistent with light source, synchronized with objects). We separate shadow from reflection because their criteria differ: reflection needs matching mirrored content, shadow directional consistency.

Full Sub-question Inventory. Table 10 lists the complete set of binary sub-questions used by the evaluator for each of the 13 laws across the three domains. Each sub-question is a yes/no observation probe, and the last column of Table 10 gives the answer that flags a violation, so the judge has a concrete checklist for each law. Human annotators do not see these sub-questions, since exposing the checklist would increase annotation burden and risk overly mechanical scoring.

A.2.3 Excluded Categories

Thermodynamics. Phase transitions (melting, boiling) and combustion do not reliably occur in short generated videos, yield too few samples, and have ambiguous visual boundaries.

Electromagnetism. Most electromagnetic phenomena produce visual effects indistinguishable from non-electromagnetic explanations. Magnetic repulsion looks identical to collision rebound, making VLM verification of the underlying mechanism impossible.

Chemistry. Oxidation-reduction and acid-base reactions (VideoScience-Bench) are too complex for video models to produce and too specialized for evaluators to judge from frames.

Waves, Energy, and Modern physics. Covered only coarsely by VideoScience-Bench. Waves need precise spatial patterns; energy is not directly observable; modern physics has no visual standard.

Holistic metrics. We exclude holistic physics scores (VideoPhy's PC, WorldModelBench's PA, PhyGenBench's binary naturalness), which give no diagnosis of *which* laws a model violates.

Table 9: Evaluation criteria: the general adherence criteria applied to every video and the physics laws by domain, each with the question shown to human annotators.

Domain	Criterion	Human Rating Question
General	Semantic Alignment (SA)	To what extent does the video content align with the text prompt?
	Physical Temporal Validity (PTV)	To what extent does the temporal sequence of physical events follow a physically plausible order?
	Object Persistence	To what extent do objects maintain consistent appearance, shape, and existence throughout the video?
Solid-Body	Gravity	To what extent do objects and liquids move in accordance with gravity?
	Inertia	To what extent do objects maintain their state of motion—remaining stationary or continuing to move—in the absence of a plausible external force?
	Momentum	To what extent are post-collision directions of motion physically plausible?
	Impenetrability	To what extent do solid objects maintain impenetrability, avoiding passage through one another?
	Collision	To what extent do objects exhibit physically plausible responses to impact, proportional to the applied force?
	Material	To what extent do objects respond in ways consistent with their apparent material properties?
Fluid	Buoyancy	To what extent do objects sink or float in a manner consistent with their apparent density?
	Displacement	When volume is added or an object is submerged, to what extent does the liquid level rise in a physically plausible manner?
	Flow dynamics	To what extent does liquid flow along surfaces, spread, and drain in a physically plausible manner?
	Boundary interaction	To what extent does liquid splash, rebound, or split in a physically plausible manner when hitting a boundary?
	Continuity	Does the liquid avoid disappearing or appearing out of nowhere?
Optical	Reflection	To what extent do reflective surfaces display content that is physically plausible given the surrounding scene?
	Shadow	To what extent are shadows physically plausible in their direction, placement, and movement relative to light sources and objects?

B Human Evaluation Design and Quality Control

B.1 Annotator Recruitment and Participation Context

The annotation study received IRB approval before data collection. Annotators were recruited at selective universities in North America. The expert group consisted of 35 graduate students who had completed university-level physics coursework; the non-expert group included more than 400 participants with at least undergraduate-level training. The task judges observable outcomes in everyday scenes, such as whether unsupported objects fall downward, whether liquids remain continuous, and whether shadows match the light source, so it does not require expert physics training.

We use a university pool instead of Amazon Mechanical Turk, whose composition, attentiveness, and non-naïveté in repeated tasks are harder to control (Paolacci et al., 2010; Chandler et al., 2014).

The annotation task was administered as an optional extra-credit activity. Participation was voluntary, and students could decline without penalty beyond not receiving the optional extra credit. Course-

Table 10: Sub-question inventory of the VLM judge. Each law is decomposed into binary observational probes; the last column gives the answer that flags a violation of that law.

Law	Sub-question	Violation if
Solid-Body Mechanics		
Gravity	Do any unsupported objects float upward or hover in mid-air?	yes
	Does any heavy object drift or float down unrealistically slowly, as if it had no weight?	yes
	Does any object fall sideways or in an unnatural direction instead of downward?	yes
Inertia	Does any object spontaneously start moving without a visible cause?	yes
	Does any moving object suddenly stop or reverse direction with no visible contact or force?	yes
Momentum	Does any object fly off in a direction completely unrelated to how it was hit or pushed?	yes
	Does the recoil direction contradict the direction of the applied force?	yes
Impenetrability	Does any object pass through another solid object as if it were not there?	yes
	Does any part of an object clip into a wall, floor, or another object's interior?	yes
Collision	Does any object remain completely unaffected after being clearly hit by another object?	yes
	Does the collision response look wildly too weak or too strong for the visible impact?	yes
	Does any object shatter or deform dramatically from a very light touch with no reasonable force?	yes
Material	Does any rigid material (glass, metal, stone) bend or stretch like rubber?	yes
	Does any soft material (cloth, rubber, rope) behave as if it were completely rigid and unbending?	yes
Fluid Dynamics		
Buoyancy	Does any heavy object (metal, stone) float on the liquid surface?	yes
	Does any light object (wood, cork, plastic) sink to the bottom?	yes
Displacement	Does the liquid level remain completely unchanged when a large object is submerged?	yes
	Does a full container spill or overflow when more liquid or an object is added?	no
	Does the liquid level behave in a clearly impossible way, such as dropping when volume is added?	yes
Flow dynamics	Does liquid flow uphill or against gravity without any force pushing it?	yes
	On a flat surface, does liquid spread outward and become thinner over time?	no
	Does liquid suddenly stop flowing or freeze in place without an obvious reason?	yes
Boundary interaction	Does liquid ignore a solid boundary and continue moving as if nothing were there?	yes
	Does liquid striking a surface produce visible droplets, spray, or ripples?	no
	Does liquid accumulate or pool on the wrong side of a barrier?	yes
Fluid continuity	Does a continuous pour or flow stay connected as a stream without sudden gaps?	no
	Does liquid disappear into nothing (not counting brief splashes)?	yes
	Does liquid appear from nowhere with no visible source?	yes
Optical Physics		
Reflection	Do reflections show completely unrelated content not present in the scene?	yes
	Does a reflection remain completely static while the scene clearly changes around it?	yes
Shadow	Do different shadows in the same scene point in contradictory directions?	yes
	Does any shadow remain fixed in place while its object clearly moves?	yes

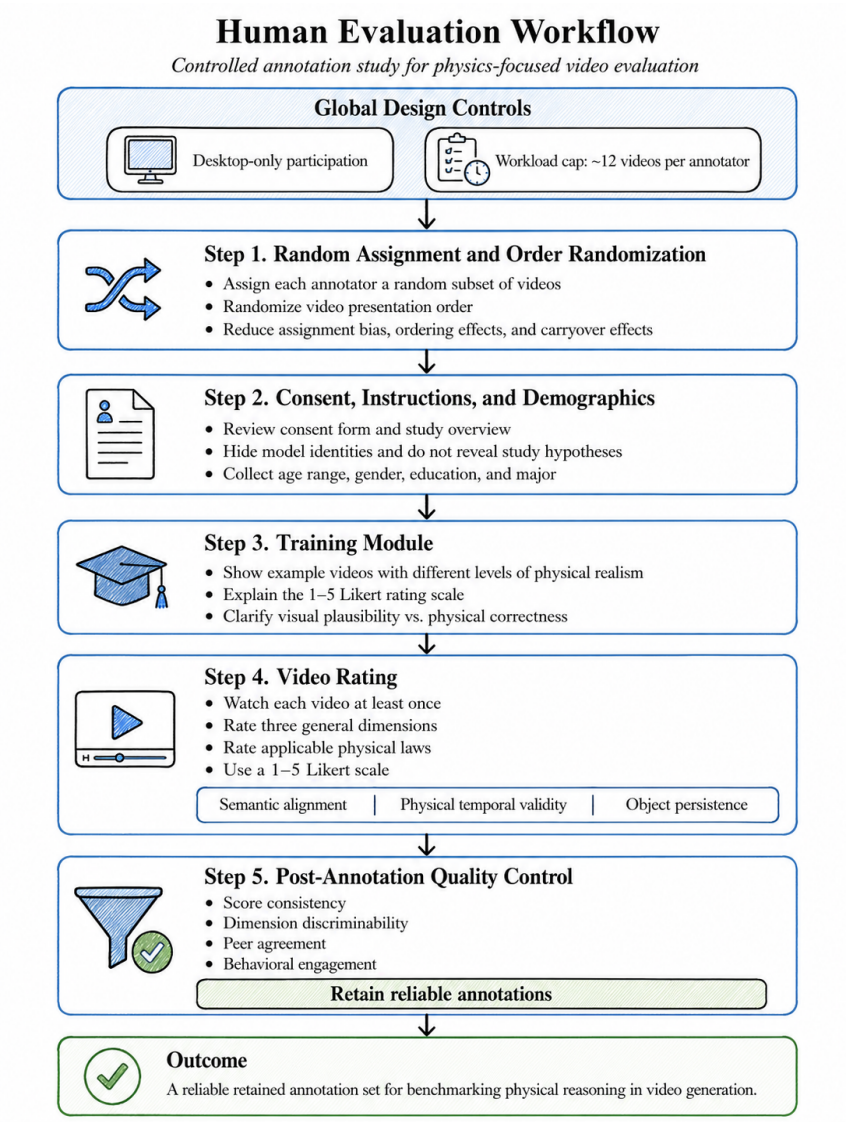


Figure 4: Annotation design workflow.

credit records were separated from research annotation records, and individual annotation content was not used to evaluate students' academic performance.

B.2 Evaluation Design

We design the annotation process as a controlled human-measurement study, following principles of valid, reliable measurement (Campbell & Stanley, 1963; Reis & Judd, 2000; Shadish et al., 2002), to reduce noise unrelated to the videos' physical quality. Figure 4 summarizes the workflow.

Viewing Environment and Workload Control. Annotators complete the task on desktop devices only, which removes variation in screen size, resolution, and interface layout. To limit fatigue and low-effort responding (Krosnick, 1991), each annotator rates roughly 12.6 videos on average.

The experimental steps are as follows:

Table 11: Vote statistics of PhyGround before annotator filtering. An *annotation* is one annotator's complete submission of all dimension scores for a single video; a *label* is one such per-dimension score, covering SA, PTV, persistence, and each physics law applicable to that video.

Basic Statistics	
# annotations	5,796
# annotators	459
# annotations per video	2.90
# labels	37.4K

1. **Random Assignment and Presentation Order.** When a subject enters our platform, they are randomly assigned a subset of videos, and the presentation order is randomized per annotator. This reduces assignment bias, ordering and carryover effects.
2. **Consent, Task Framing, and Demographic Information.** Participants first review the consent form, study overview, voluntary-participation statement, and other required information. The task is described without revealing our hypotheses, and model identities are hidden from annotators, to reduce demand effects (Orne, 1962). After consenting, participants report age range, gender, educational background, and academic major, which are used only for aggregate characterization of the annotator pool.
3. **Training Module.** Annotators then complete a short training module (Figure 5) that walks through three worked examples at different levels of physical realism, with per-dimension scores and reasoning shown for each, and explains how minor imperfections, moderate inconsistencies, and severe violations map onto the five-point Likert scale.
4. **Rating Procedure.** Each video must be watched at least once before rating. Annotators rate the three general dimensions (semantic alignment, physical temporal validity, object persistence) and the 2–4 physical laws applicable to that video, each on a 1–5 scale. Each video is rated by multiple annotators, and the final score is the mean across annotators, the per-video aggregation rule used by the leaderboard.

Quality control (Appendix B.3) then removes inattentive or non-discriminating annotators.

Table 11 gives pre-QC annotation, annotator, and label counts; each video has at least two annotators.

B.3 Quality Control Pipeline

The raw annotation data went through two rounds of quality control based on four signals:

1. **Score consistency**: An annotator's score standard deviation. std $= 0$ (one score for every dimension of every video) carries no information; std < 0.3 counts as near-constant.
2. **Per-video Copy-paste Rate**: Whether the annotator gives identical scores to all dimensions within each video. A 100% copy-paste rate means the annotator does not differentiate evaluation dimensions (e.g., gravity vs. collision), so the labels are uninformative for per-law scoring even when the annotator's cross-video std is greater than zero.
3. **Peer MAE**: The mean absolute deviation between an annotator's scores and those of other annotators on the same video-dimension cells, averaged over the annotator's pairwise comparisons. On the 1–5 scale, peer MAE above 1.8 is flagged as extreme; this is roughly twice the population's typical pairwise disagreement. Because honest annotators occasionally disagree (e.g., stricter or more lenient calibration), peer MAE alone is not a removal criterion; it must co-occur with a behavioral anomaly or a high copy-paste rate.
4. **Behavioral Signals**: Two platform-side signals are recorded for every submission. *Page stay time* is the elapsed seconds between page load and rating submission; a median below 30 seconds across an annotator's submissions is treated as short stay, and below 10 seconds as too short for any deliberate judgment. *Video play count* is the number of times the annotator triggered playback before submitting; a maximum play count of zero across all submissions means the rating was entered without ever playing the video. Either signal in isolation is

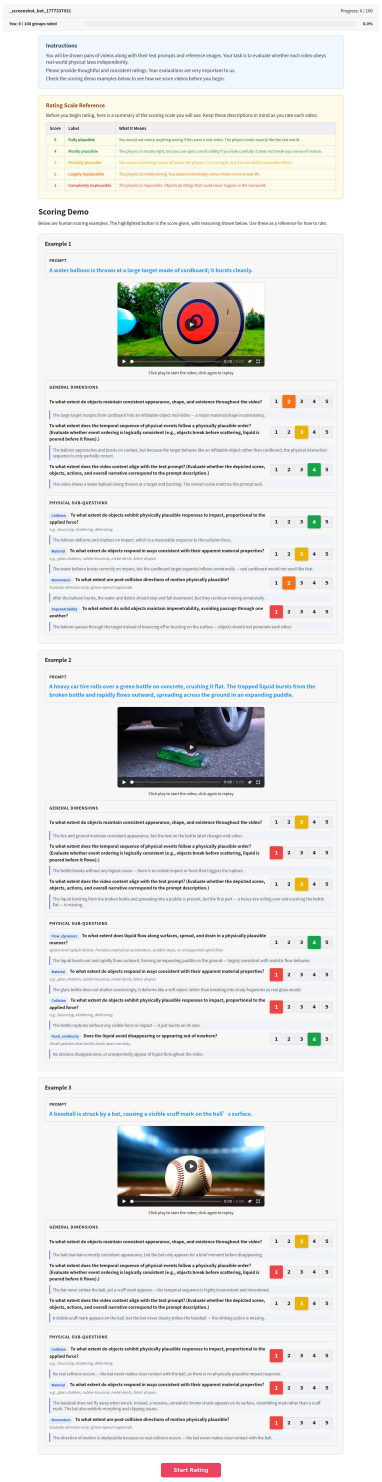


Figure 5: Training module shown to annotators before rating: the rating scale and three worked examples (a water balloon hitting a cardboard target, a tire crushing a glass bottle, a baseball struck by a bat), each with its per-dimension scores highlighted and a short justification.

noisy, so a behavioral anomaly removes an annotator only when both signals agree (no video ever played and short stay) or together with copy-paste or peer-MAE evidence.

The largest source of removals is annotators who never played any video and had a short median stay; constant or near-constant scoring is the second. The remaining removals require several signals to be jointly abnormal: a copy-paste rate $\geq 75\%$ combined with short stay time or unplayed videos, extremely low annotation volume or stay time, or extreme peer MAE paired with a behavioral anomaly. In total, 23.4% of annotators and their ratings were removed, leaving 352 annotators.

B.4 Difficult-Item Analysis

To check whether hard items lead to skipped or careless judgments, we bucket every video with at least two post-QC annotators by its inter-annotator pairwise MAE (Table 12). Skip rates are low in every bucket and do not rise with difficulty. Mean stay time rises monotonically with difficulty, so annotators spend more time on harder items. The copy-paste rate is highest on easy items, where the dimensions genuinely deserve similar scores, and rises again on the very-hard bucket, where some annotators fall back to a single score when uncertain. These annotations are a minority and are diluted by the per-video mean. Pairwise MAE is similar across the general dimensions and the physics domains (Table 13), so no dimension is markedly harder to annotate than the others.

Table 12: Annotator behaviour by item difficulty on the post-QC pool. Difficulty is the video's inter-annotator pairwise MAE on the 1–5 scale (easy: 0–0.75, medium: 0.75–1, hard: 1–1.5, very hard: $\geq$1.5). Skip rate is the share of assignments the annotator skipped; stay is the mean page stay time in seconds; copy-paste is the share of annotations that gave every dimension the same score. Quality control removes annotators on scoring and behavioral signals; peer MAE enters only jointly with a behavioral anomaly.

Difficulty	Videos	Skip (%)	Stay (s)	Copy-paste (%)
easy	330	1.9	274	21.9
medium	238	1.0	332	10.2
hard	708	1.8	368	10.4
very hard	555	1.6	798	17.7

Table 13: Inter-annotator pairwise MAE per general dimension and per physics domain (laws pooled), post-QC pool.

Dimension	Pairwise MAE
SA	1.21
PTV	1.24
Persist.	1.23
Solid-Body	1.25
Fluid	1.29
Optical	1.23

B.5 Ranking Robustness

The aggregate model ranking is reproducible across complementary resampling protocols. Table 14 reports, per model, the mean-aggregated Overall score (the leaderboard aggregation), two 95% bootstrap confidence intervals (CI), and a median-aggregation comparison. The prompt CI resamples prompts, probing sensitivity to the sampled prompt set, while the annotator CI resamples annotator scores within each fixed video, probing sensitivity to which annotators rated each video.

The mean-vs.-median comparison is a stronger perturbation because it changes the aggregation rule itself; even so, none of the eight model ranks move.

Split-half rank-correlation intervals. In 100 random trials, the 352 post-QC annotators are split into two equal halves; every video model is scored in each half with the leaderboard aggregation, and the two eight-model score vectors are compared with Spearman ρ and Kendall τ-b (Table 3). Even the lower ends of the 95% intervals over the 100 splits remain large (Spearman $\rho \geq 0.773$, Kendall τ-b ≥ 0.605): independently sampled annotator halves recover the same model ordering.

Annotator-pool sub-sampling. We probe how the leaderboard responds to shrinking the annotator pool. For each setting, we draw 200 random annotator sub-samples and recompute every model's mean over the general dimensions, then report the root-mean-square deviation (RMS Δ, 1–5 units) from the full-pool baseline (Table 15). Dropping a random 30% of annotators barely moves per-model

Table 14: Leaderboard robustness under the mean-aggregation pipeline. Prompt CI resamples prompts with replacement; annotator CI resamples annotator scores within each video; the last two columns replace the per-video annotator mean with the annotator median.

Rank	Model	Mean score	Prompt 95% CI	Annotator 95% CI	Median rank	Median–mean
1	Wan2.2-27B-A14B	3.282	3.197–3.371	3.234–3.343	1	+0.019
2	Veo-3.1	3.280	3.184–3.373	3.228–3.339	2	+0.006
3	OmniWeaving	3.101	3.021–3.185	3.060–3.156	3	+0.007
4	Cosmos-14B	2.904	2.824–3.009	2.852–2.947	4	+0.004
5	LTX-2.3-22B	2.698	2.600–2.785	2.644–2.743	5	-0.012
6	Wan2.2-TI2V-5B	2.628	2.528–2.708	2.585–2.682	6	+0.001
7	Cosmos-2B	2.585	2.488–2.697	2.546–2.629	7	+0.010
8	LTX-2-19B	2.564	2.475–2.645	2.520–2.611	8	-0.021

scores, while sub-sampling to 20 annotators inflates the deviation substantially; the observed QC drop moves scores more than a random drop of the same size.

Table 15: Annotator-pool sub-sampling stability of per-model leaderboard scores. RMS Δ is the root-mean-square deviation of a model's mean over the general dimensions from its full-pool post-QC value, in score units; mean and max are taken across the leaderboard models. Random rows average over repeated sub-samples; the observed QC drop compares the pre-QC pool with the post-QC pool.

Sub-sample	mean RMS Δ	max RMS Δ
Drop a random fraction of annotators		
drop 5 %	0.0115	0.0122
drop 10 %	0.0172	0.0197
drop 20 %	0.0247	0.0290
drop 30 %	0.0333	0.0375
Observed quality-control drop		
pre-QC pool $\rightarrow$ post-QC pool	0.0585	0.1065
Keep only N uniformly sampled annotators		
$N = 10$	0.2944	0.3209
$N = 20$	0.2037	0.2391
$N = 50$	0.1228	0.1458
$N = 100$	0.0794	0.0894

B.6 Annotator Diversity

To quantify how broadly annotators spread across demographic facets, we compute an Annotator Diversity Index (ADI): for each facet (gender, education, major, age, cohort) we report the effective number of categories $\exp(H)$, with H the Shannon entropy of annotators over the facet's categories, and take the geometric mean over the five facets as the ADI. Table 16 compares the per-facet values with a hypothetical expert panel: twelve CS PhD students from one lab, gender-balanced and aged 26–40. Our ADI is substantially higher, with the largest gaps on *major* and *cohort*.

C Aggregation Sensitivity

The Physics score averages each model's (video, law) units within each domain and weights domains by their number of units. We test this choice by recomputing Overall scores with equal-law aggregation schemes that remove prompt-count weighting.

C.1 Schemes

- **Sample-weighted** (the leaderboard rule): the domain mean is the mean of all (video, law) units in that domain, so laws with more prompts pull harder; the Physics aggregate is then $(693 \cdot \text{Solid} + 136 \cdot \text{Fluid} + 47 \cdot \text{Optical})/876$.

Table 16: Annotator Diversity Index (ADI). Per demographic facet, the effective number of categories $\exp(H)$ of the annotator distribution; buckets counts the facet's coded categories. The expert column is a hypothetical single-lab CS PhD panel. ADI is the geometric mean of $\exp(H)$ over the facets.

Facet	Ours $\exp(H)$	Buckets	Expert $\exp(H)$
Gender	2.15	4	2.00
Education	1.90	3	1.00
Major	3.17	8	1.00
Age	3.47	6	2.00
Cohort	2.95	4	1.00
ADI	2.66		1.32

- **Equal-law-within-domain**: the domain mean is the unweighted average of the per-law means inside that domain, so within Solid-Body the six laws each contribute $1/6$ regardless of prompt count, while Physics still weights domains by their unit counts.
- **Equal-law-global**: the Physics aggregate is the unweighted mean of all 13 per-law means, ignoring domain grouping; each law contributes $1/13$ regardless of prompt count or domain size, so the implied domain weights become $6/13$, $5/13$ and $2/13$.

General averages SA, PTV and Persistence unweighted, and Overall $= 0.5 \cdot$ General $+ 0.5 \cdot$ Physics.

C.2 Per-Domain Δ from Removing Within-Domain Law-Count Imbalance

Table 17 reports, per domain, the absolute gap between the equal-law-within-domain domain mean and the sample-weighted domain mean, summarized across the eight models.

Table 17: Per-domain gap Δ between the equal-law-within-domain and the sample-weighted domain mean, in absolute value, summarized across the comparison models.

Domain	Mean abs Δ	Max abs Δ	Model with max abs Δ
Solid-Body	0.014	0.03	Cosmos-2B / Veo-3.1
Fluid	0.014	0.03	Cosmos-14B
Optical	0.030	0.07	Cosmos-2B

C.3 Overall Sensitivity Under Each Alternative Scheme

Table 18 reports the absolute Overall shift and rank changes under each alternative scheme.

Table 18: Sensitivity of the Overall score and ranking to the alternative Physics aggregation schemes; shifts and rank changes are relative to the sample-weighted baseline.

Scheme	Max abs Δ	Mean abs Δ	Rank changes	Max rank shift
Equal-law-within-domain	0.015	0.007	2/8	1
Equal-law-global	0.099	0.063	2/8	1

C.4 Takeaway

Within-domain law imbalance contributes essentially nothing to the Overall score. Equal-law-global shifts absolute Overall scores more, but no model moves by more than one rank. The only persistent swap is the Wan2.2-27B-A14B $\leftrightarrow$ Veo-3.1 exchange at the top of the leaderboard, which is already a near-tie under sample-weighting.

Table 19: Per-law human evaluation breakdown on PhyGround (1–5 scale, higher is better). Each column reports the per-model mean over all videos annotated for that law (or merged group). Laws with too few prompts per model are merged with a semantically and empirically related law: Liq-Solid (liquid–solid interaction) = boundary interaction + buoyancy; Conserv. (fluid mass/volume conservation) = displacement + fluid continuity; Optical = reflection + shadow. **Bold** marks the best score in each column and underline marks the second-best. †: closed-source.

	Solid-Body						Fluid			Optical
Model	Gravity	Inertia	Momentum	Impen.	Collision	Material	Liq-Solid	FlowDyn	Conserv.	Optical
Wan2.2-27B-A14B	**3.23**	3.23	**3.15**	**3.49**	**3.13**	3.17	3.15	3.17	3.21	3.55
Veo-3.1†	3.21	**3.32**	2.94	3.30	2.99	3.14	**3.70**	**3.82**	**3.51**	**3.69**
OmniWeaving	2.97	2.90	2.89	3.15	2.82	**3.18**	3.05	3.39	3.38	3.22
Cosmos-14B	2.77	2.83	2.75	3.05	2.69	2.82	2.92	2.82	3.14	3.36
LTX-2.3-22B	2.62	2.56	2.54	2.66	2.55	2.77	3.02	2.93	3.18	2.98
Wan2.2-TI2V-5B	2.66	2.58	2.46	2.75	2.47	2.59	2.71	2.62	2.76	2.98
Cosmos-2B	2.50	2.75	2.28	2.87	2.36	2.59	2.62	2.86	2.84	3.35
LTX-2-19B	2.55	2.62	2.35	2.45	2.42	2.49	2.72	3.23	3.23	2.78

D Per-Law Score Breakdown

Table 19 reports the human-evaluation mean for every individual law.

Observation 1: Within solid-body, the per-law leaderboard does not match the domain leaderboard. Even though Wan2.2-27B-A14B leads the solid-body domain aggregate, two of the six solid-body laws (Inertia, Material) are won by other models, and Material is the only law in the table where neither the leading open-source model nor the closed-source model is best.

Observation 2: The solid-body bottleneck is concentrated in Momentum and Collision. Aggregating the table column-wise, Momentum and Collision are the two hardest laws across the entire table, both noticeably below Impenetrability (the easiest solid-body axis) and well below Optical. What separates the easy and hard solid-body laws is whether contact dynamics are involved: free-motion laws (Gravity, Inertia) and the static non-overlap constraint (Impenetrability) score uniformly higher than the dynamic-contact laws (Momentum, Collision).

Observation 3: Veo-3.1's fluid advantage lies in liquid motion and liquid–solid interaction. Veo-3.1 leads all three fluid columns of the table, but by very different margins: the lead is decisive on liquid–solid interaction (boundary interaction and buoyancy) and on flow dynamics, while on conservation (displacement and continuity) the lead over OmniWeaving collapses to a near-tie.

Observation 4: Per-law within-model spread separates balanced from spiky physics profiles. The spread between a model's best and worst law differs substantially across models: Wan2.2-27B-A14B is the most balanced, with a small spread and no obviously weak law. Cosmos-2B is the most spiky: its spread crosses more than a full rating level between its weakest and strongest laws. LTX-2-19B flips rank the most: last on Impenetrability, competitive on the fluid laws.

E Per-Law Qualitative Examples of Judge–Human Alignment

Figures 6 to 18 show one qualitative case per physical law. Each case is selected so that PhyJudge-9B closely tracks the human rating on the focal law while Gemini-3.1-Pro disagrees by at least two rating levels, and so that PhyJudge-9B is also the closer of the two judges on the overall score. For every case we show five evenly spaced frames from the generated clip, the prompt, the per-dimension score table comparing Human, PhyJudge-9B, and Gemini-3.1-Pro, and the verbatim conclusion sentence Gemini emits about the focal law. In the score tables, a green cell marks a dimension on which PhyJudge-9B is within one rating level of the human score and closer to it than Gemini, and a red cell marks a dimension on which Gemini-3.1-Pro is at least two levels from the human score. Both judges use the same sub-question prompt schema as in the main results: Gemini-3.1-Pro with chain-of-thought (CoT) reasoning on top of the sub-questions, and PhyJudge-9B with the score-only sub-question prompt that matches its training format.



Domain: Solid-Body · Collision Model: Cosmos-14B
Prompt: A steel ball bearing rolls across a flat surface and collides with another ball bearing, both ceasing motion.

Judge	SA	PTV	Persist.	Collision	Impenetr.	Material	Momentum	Overall
Human	4.3	4.7	4.3	4.7	5.0	5.0	4.3	4.6
PhyJudge	5	4	4	5	3	4	3	4.0
Gemini	1	1	1	1	5	1	1	1.5

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Collision:
"The collision violates basic physics such as conservation of momentum and features an unrealistic transformation"

Figure 6: **Collision** on a Cosmos-14B generation: two steel ball bearings collide and come to rest. Humans rate the collision realistic and PhyJudge-9B agrees, while Gemini-3.1-Pro reads it as a near-total violation: a contact-resolution case where it over-penalizes plausible post-impact rest.



Domain: Solid-Body · Gravity Model: LTX-2.3-22B
Prompt: A person throws a frisbee that tumbles along the ground before stopping.

Judge	SA	PTV	Persist.	Collision	Gravity	Impenetr.	Momentum	Overall
Human	3.0	2.0	1.0	1.5	1.0	2.5	1.0	1.8
PhyJudge	3	3	2	2	1	3	1	2.2
Gemini	3	5	5	5	5	5	5	4.7

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Gravity:
"The video shows a frisbee being thrown and falling to the ground in a manner consistent with gravity. No unnatural floating or falling is observed"

Figure 7: **Gravity** on an LTX-2.3-22B generation: a thrown frisbee tumbles along the ground. Humans rate gravity broken and PhyJudge-9B agrees, while Gemini-3.1-Pro gives the top score, missing an obvious free-fall failure that humans flag.



Domain: Solid-Body · Impenetr. Model: OmniWeaving
Prompt: A motorcycle rests on its kickstand on dry, textured sand. As the bike is lifted and begins to roll forward, the sharp, dark shadows cast by its front wheel, intricate brake rotor, and suspension forks stretch and glide dynamically across the uneven ground.

Judge	SA	PTV	Persist.	Impenetr.	Inertia	Material	Shadow	Overall
Human	4.5	4.5	5.0	5.0	5.0	5.0	4.0	4.7
PhyJudge	5	5	4	5	5	4	5	4.7
Gemini	2	1	1	1	5	1	5	2.2

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Impenetr.:
"The video shows severe clipping and morphing of solid objects, violating impenetrability"

Figure 8: **Impenetrability** on an OmniWeaving generation: a motorcycle is lifted off its kickstand and begins to roll on textured sand. Humans see no inter-penetration and PhyJudge-9B agrees, while Gemini-3.1-Pro gives the lowest score despite the absence of any visible solid–solid overlap.



Domain: Solid-Body · Inertia **Model: Veo-3.1**
Prompt: The grasshopper jumps from the ground into the air amidst the surrounding vegetation. It reaches a peak height and then falls back down to the ground.

Judge	SA	PTV	Persist.	Gravity	Impenetr.	Inertia	Overall
Human	4.5	4.0	3.5	4.5	4.5	4.0	4.2
PhyJudge	2	5	4	5	4	4	4.0
Gemini	4	2	1	1	1	1	1.7

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Inertia:
"The grasshopper's sudden appearance in mid-air and its unnatural, non-parabolic trajectory severely violate the principles of inertia"

Figure 9: **Inertia** on a Veo-3.1 generation: a grasshopper jumps from the ground, reaches a peak, and falls back. Humans rate the inertial trajectory plausible and PhyJudge-9B agrees, while Gemini-3.1-Pro treats the same ballistic motion as a near-total inertia violation.



Domain: Solid-Body · Material **Model: Wan2.2-27B-A14B**
Prompt: A person throws a throwing axe at a large pumpkin; the axe cuts cleanly through the pumpkin.

Judge	SA	PTV	Persist.	Collision	Gravity	Material	Momentum	Overall
Human	4.3	4.7	4.7	4.7	4.7	4.3	4.7	4.6
PhyJudge	4	3	3	4	3	4	3	3.4
Gemini	2	2	1	1	1	1	1	1.3

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Material:
"The rigid axe completely loses its structural integrity and transforms into an unrecognizable soft material, which is a severe violation of material properties"

Figure 10: **Material** on a Wan2.2-27B-A14B generation: a thrown axe cuts cleanly through a pumpkin. Humans rate the material response plausible and PhyJudge-9B tracks them, while Gemini-3.1-Pro gives the lowest score on a case where humans accept the cut behavior.



Domain: Solid-Body · Momentum **Model: LTX-2-19B**
Prompt: A player kicks a kickball that bounces off a wooden fence, changing direction.

Judge	SA	PTV	Persist.	Collision	Impenetr.	Material	Momentum	Overall
Human	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
PhyJudge	1	1	1	3	1	1	1	1.2
Gemini	5	5	1	1	5	5	5	3.8

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Momentum:
"The direction of motion after the kick and the bounce off the fence are physically realistic and consistent with the law of momentum"

Figure 11: **Momentum** on an LTX-2-19B generation: a player is supposed to kick a kickball that bounces off a wooden fence, but the generated clip contains no kick and no bounce, so no momentum transfer is depicted. Humans rate momentum broken and PhyJudge-9B agrees, while Gemini-3.1-Pro reports near-perfect momentum behavior on the same clip.



Domain: Fluid · Boundary **Model: LTX-2.3-22B**
Prompt: A golden-colored liquid pours from a transparent, narrow-necked bottle into a clear glass filled with ice cubes. The liquid flows in a steady, continuous stream, cascading over the ice and gradually filling the glass.

Judge	SA	PTV	Persist.	Boundary	Displace.	Flow	Fluid Cont.	Overall
Human	4.7	4.3	4.3	4.7	4.7	4.7	4.7	4.6
PhyJudge	5	3	5	4	3	3	2	3.7
Gemini	3	2	1	1	1	4	1	1.9

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Boundary:
"The liquid stream pours into the glass and hits the surface of the liquid and ice, but there is absolutely no interaction. The liquid in the glass remains completely frozen in time, with no ripples, splashes, or displacement, and the pouring stream simply disappears at the contact point. This is a severe failure of physical realism regarding boundary interaction"

Figure 12: **Boundary interaction** on an LTX-2.3-22B generation: a golden liquid pours over ice cubes inside a glass. Humans see realistic deflection and splash on the ice and PhyJudge-9B follows them, while Gemini-3.1-Pro reads the same liquid–solid interaction as broken: a representative fluid case where Gemini-3.1-Pro is far more pessimistic than human raters.



Domain: Fluid · Buoyancy **Model: Cosmos-2B**
Prompt: A ski jet is being towed by a larger boat on calmer waters.

Judge	SA	PTV	Persist.	Buoyancy	Flow	Inertia	Reflection	Overall
Human	3.5	4.5	5.0	4.5	4.5	4.5	4.5	4.4
PhyJudge	3	4	5	5	3	5	4	4.1
Gemini	1	5	5	1	1	1	4	2.7

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Buoyancy:
"The heavy boat hovering in the air rather than resting in the water is a severe violation of physical laws"

Figure 13: **Buoyancy** on a Cosmos-2B generation: a jet ski is towed across water. Humans rate buoyancy plausible and PhyJudge-9B agrees, while Gemini-3.1-Pro flags the floating motion.



Domain: Fluid · Displace. **Model: LTX-2.3-22B**
Prompt: Water is dispensed into plastic bottles on a rotary filling machine. The bottles are filled with water.

Judge	SA	PTV	Persist.	Boundary	Displace.	Flow	Fluid Cont.	Overall
Human	5.0	4.5	5.0	4.5	5.0	5.0	5.0	4.9
PhyJudge	5	2	5	3	5	3	5	4.0
Gemini	5	1	1	2	1	1	1	1.8

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Displace.:
"The liquid volume behaves in a severely incorrect and impossible manner"

Figure 14: **Displacement** on an LTX-2.3-22B generation: water is dispensed into plastic bottles on a rotary filling machine, with sustained flow and steady volume transfer into each bottle. Humans rate the volume transfer plausible and PhyJudge-9B matches them, while Gemini-3.1-Pro reads continuous filling as a violation, an instance of its under-scoring of fluid laws.



Domain: Fluid · Flow **Model: Veo-3.1**
Prompt: A chef in a modern kitchen tips a white bowl, pouring its contents in a steady stream that flows smoothly down into a pot on the stove. He then stirs the mixture with a spoon, causing the liquid to swirl and flow in continuous, circular currents.

Judge	SA	PTV	Persist.	Boundary	Displace.	Flow	Fluid Cont.	Overall
Human	5.0	5.0	4.5	4.0	3.5	5.0	4.5	4.5
PhyJudge	3	4	2	1	4	5	4	3.2
Gemini	4	3	2	1	2	1	1	2.1

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Flow:
"The liquid stream during the pour looks like a solid, static cylinder rather than a flowing fluid. Additionally, the bowl remains completely full after a large amount of liquid is supposedly poured out, showing a severe lack of realistic fluid dynamics"

Figure 15: **Flow dynamics** on a Veo-3.1 generation: a chef pours from a bowl into a pot and stirs the resulting mixture. Humans rate bulk-flow patterns realistic and PhyJudge-9B matches them, while Gemini-3.1-Pro gives the lowest score on the same clip.

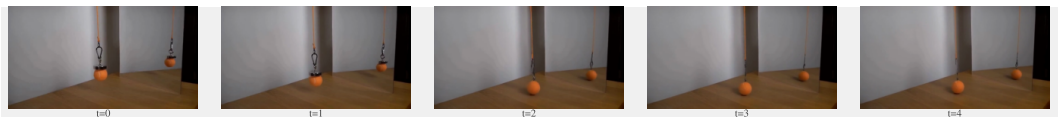


Domain: Fluid · Fluid Cont. **Model: Veo-3.1**
Prompt: A worker opens the lid of a large stainless steel tank and pours in a white bucket of ingredients. The ingredients fall into the tank and mix with its contents.

Judge	SA	PTV	Persist.	Displace.	Flow	Fluid Cont.	Reflection	Overall
Human	1.5	1.5	1.5	2.5	2.0	2.5	2.0	1.9
PhyJudge	5	2	4	2	3	3	4	3.3
Gemini	4	5	5	5	5	5	5	4.8

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Fluid Cont.:
"The fluid continuity is perfectly maintained throughout the pouring action. The stream remains connected, and no liquid inexplicably appears or disappears"

Figure 16: **Fluid continuity** on a Veo-3.1 generation: a worker pours a bucket of ingredients into a stainless tank. Humans see partial continuity violations and PhyJudge-9B agrees, while Gemini-3.1-Pro gives the top score, the opposite of its usual under-scoring of fluid laws.



Domain: Optical · Reflection **Model: LTX-2-19B**
Prompt: A tennis ball attached to a magnet and string is hanging in front of a mirror and creating an illusion of two tennis balls. The string lowers the ball slowly. The illusion of two tennis balls moves downwards as the real ball is lowered. Static shot with no camera movement.

Judge	SA	PTV	Persist.	Gravity	Reflection	Overall
Human	2.5	3.0	1.5	3.0	3.0	2.7
PhyJudge	3	4	2	4	3	3.2
Gemini	4	5	5	5	5	4.8

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Reflection:
"The reflection accurately mirrors the objects, colors, and movements in the scene without any unrelated content or static behavior"

Figure 17: **Reflection** on an LTX-2-19B generation: a tennis ball is supposed to be lowered in front of a mirror, but the generated object morphs through several distinct shapes during the clip. Humans penalize this failure and PhyJudge-9B matches them, whereas Gemini-3.1-Pro assigns the top score, failing to notice that the reflected object changes identity across frames.



Domain: Optical · Shadow **Model: LTX-2-19B**
Prompt: The person in blue launches the yellow frisbee across the lawn. The frisbee flies through the air, following a parabolic trajectory before landing on the grass.

Judge	SA	PTV	Persist.	Collision	Impenetr.	Inertia	Shadow	Overall
Human	1.5	1.0	1.0	1.0	1.0	1.5	1.0	1.1
PhyJudge	3	3	1	2	1	2	2	2.0
Gemini	2	5	1	1	1	5	5	2.8

Gemini-3.1-Pro reasoning (verbatim, conclusion sentence):
Shadow:
"The shadows in the video are consistent with the lighting conditions and object movements, showing no physical inaccuracies"

Figure 18: **Shadow** on an LTX-2-19B generation: a yellow frisbee flies across a lawn and lands on the grass. Humans rate the cast shadow inconsistent with the frisbee's trajectory and PhyJudge-9B is close, while Gemini-3.1-Pro gives the top score, missing the optical mismatch that humans see.

F Judge Bias without Per-Law Sub-Questions

We ablate per-law sub-questions by training an *Ours (no SubQ)* 9B checkpoint on the same data but with a prompt that omits the sub-questions and asks for one dimension's score (Table 20). Against the PhyJudge-9B row of Table 5, removing the sub-questions raises bias on every column.

Table 20: Aggregate relative bias (%) with the baseline prompt that carries no per-law sub-questions. General, Physics and Overall are the absolute grouped mean-bias $E = |\bar{J} - \bar{H}|/\bar{H}$, macro-averaged; Signed is the signed relative bias of the per-video overall score. **Bold** marks the best value per column. *Ours (no SubQ)* is the 9B checkpoint trained with the same no-sub-question template.

Model	**General (%)** $\downarrow$	**Physics (%)** $\downarrow$	**Overall (%)** $\downarrow$	**Signed (%)** $\rightarrow 0$
Claude (Opus 4.7)	19.7	**7.8**	13.7	+10.7
Gemini-3.1-Pro	20.9	37.1	29.0	-22.7
Qwen9B-base	35.8	29.8	32.8	+38.5
Qwen27B-base	36.8	36.2	36.5	+40.4
Ours (no SubQ)	**7.5**	10.2	**8.9**	**-6.6**

G Per-Law Support, Confidence Intervals, and Split-Half Stability

The prompt suite is imbalanced across laws: solid-body laws carry several times more prompts per model than fluid or optical laws (Table 8). We report, per law, a bootstrap confidence interval on the cross-model mean and the split-half stability of the 8-model ranking.

Why the suite is imbalanced. The composition of the suite is determined by the filtering criteria of Appendix A.1, after de-duplication (multi-view copies of one scene collapse to a single view) and removal of prompts that even the smallest generators solve; screening a further 12k-caption corpus (OpenVid) for the under-covered laws yielded only 35 prompts that met the same criteria.

Protocol. The unit is one model's (video, law) score: the mean across annotators of that video's rating on that law, from the post-QC pool, aggregated as in the leaderboard. Confidence intervals are percentile 95% intervals from 2,000 bootstrap resamples over units. Split-half stability follows the protocol of Table 3, applied per law: over the 100 annotator-pool partitions we report the mean and standard deviation of the Spearman ρ between the two halves' 8-model vectors.

Reading Table 21. Intervals are narrowest on the well-populated solid-body laws and widest on the thinnest laws (reflection, buoyancy), and each domain aggregate is tighter than any of its laws.

Table 21: Per-law support across the evaluated models. Units are (model, video, law) cells pooled over the models; the cross-model mean carries a percentile bootstrap interval over units. Split-half ρ is the Spearman correlation between the model rankings obtained from two disjoint random halves of the annotator pool, mean $\pm$ sd over partitions. Bold rows pool each domain's laws.

Law	Mean [95% CI]	Split-half ρ
Collision	2.68 [2.63, 2.74]	0.90 ± 0.08
Gravity	2.81 [2.75, 2.87]	0.84 ± 0.10
Momentum	2.67 [2.60, 2.73]	0.90 ± 0.06
Impenetrability	2.96 [2.90, 3.03]	0.92 ± 0.05
Material	2.84 [2.78, 2.91]	0.85 ± 0.06
Inertia	2.85 [2.76, 2.94]	0.72 ± 0.15
Solid-Body	**2.79 [2.76, 2.82]**	$\mathbf{0.94 \pm 0.05}$
Boundary interaction	2.91 [2.79, 3.03]	0.50 ± 0.22
Flow dynamics	3.11 [2.99, 3.22]	0.72 ± 0.19
Fluid continuity	3.19 [3.06, 3.31]	0.38 ± 0.25
Displacement	3.12 [3.00, 3.25]	0.49 ± 0.26
Buoyancy	3.13 [2.96, 3.28]	0.67 ± 0.18
Fluid	**3.08 [3.03, 3.14]**	$\mathbf{0.72 \pm 0.14}$
Shadow	3.17 [3.06, 3.29]	0.68 ± 0.19
Reflection	3.36 [3.20, 3.51]	0.70 ± 0.17
Optical	**3.24 [3.15, 3.34]**	$\mathbf{0.79 \pm 0.14}$

Table 22: Per-model means with bootstrap intervals on reflection, buoyancy and shadow, the three thinnest laws, which the per-law breakdown reports only after merging each with a related law.

Model	Reflection	Buoyancy	Shadow
Wan2.2-27B-A14B	3.73 [3.26, 4.12]	3.37 [3.07, 3.66]	3.45 [3.05, 3.81]
Veo-3.1	3.89 [3.45, 4.28]	3.82 [3.44, 4.18]	3.57 [3.26, 3.85]
OmniWeaving	3.39 [3.07, 3.74]	3.22 [2.80, 3.64]	3.13 [2.81, 3.43]
Cosmos-14B	3.38 [2.94, 3.79]	3.18 [2.83, 3.52]	3.37 [3.07, 3.66]
LTX-2.3-22B	3.06 [2.66, 3.46]	3.20 [2.71, 3.64]	2.94 [2.64, 3.21]
Wan2.2-TI2V-5B	2.91 [2.59, 3.24]	2.77 [2.31, 3.25]	3.04 [2.78, 3.31]
Cosmos-2B	3.66 [3.14, 4.12]	2.81 [2.31, 3.25]	3.17 [2.81, 3.52]
LTX-2-19B	2.86 [2.46, 3.23]	2.64 [2.22, 3.06]	2.73 [2.39, 3.07]

Table 23: Bootstrap intervals for the per-domain contrasts of the main results. Each model's (video, law) units in the domain are resampled independently; Δ is the difference of domain means.

Contrast (A $-$ B)	Domain	Δ [95% CI]
Wan2.2-27B-A14B $-$ Veo-3.1	Solid-Body	0.11 [0.01, 0.20]
Veo-3.1 $-$ Wan2.2-27B-A14B	Fluid	0.47 [0.27, 0.68]
Cosmos-2B $-$ LTX-2-19B	Optical	0.57 [0.17, 0.97]
LTX-2-19B $-$ Cosmos-2B	Fluid	0.27 [0.06, 0.48]

Ranking stability across annotator halves follows the same order: the large solid-body laws and the Solid-Body aggregate rank the eight models stably, the Fluid and Optical aggregates less so, and fluid continuity, displacement and boundary interaction are the least stable laws. Table 19 merges every law with fewer than 30 prompts per model into a related column; the per-model means of reflection, buoyancy and shadow with intervals are given in Table 22.

Headline gaps. Table 23 gives bootstrap CIs for the four per-domain contrasts used in Section 4.2.

H VLM JUDGE TRAINING DETAILS

Table 24 lists the judge fine-tune's LoRA, optimization, and compute configuration. Checkpoints are selected on a random 5% of the training videos from the seven training generators.

Table 24: LoRA, optimization, and compute configuration of the judge fine-tune.

Group	Parameter	Value
LoRA	Adapter rank / alpha / dropout	32 / 64 / 0.05
	Target modules	all linear layers (q, k, v, o, MLP gate/up/down)
	ViT	frozen
	Vision–language aligner	trainable, learning rate 2×10^{-6}
Optimization	Optimizer	AdamW (β_1=0.9, β_2=0.95, weight decay 0.1)
	Learning rate	1×10^{-4} (LoRA), 2×10^{-6} (aligner)
	Schedule	cosine decay with 5% linear warmup
	Gradient clipping	max-norm 1.0
Batching	Per-device batch size	1
	Gradient accumulation	8 steps
	Epochs	1
	Maximum sequence length	8192 tokens
Systems	Distributed strategy	DeepSpeed ZeRO-2
	Numerical precision	bfloat16
Selection	Validation	every 100 optimizer steps; best checkpoint by validation loss
Compute	Hardware	4× NVIDIA A100-SXM4-80 GB
	Peak memory	∼48 GiB / GPU
	Wall-clock	∼1 h 42 min (294 optimizer steps)
	Throughput	1.53 samples/s

H.1 TRAINING DATA

Each training sample contains three elements:

- **Video frames**: sampled at 4fps (consistent with inference), each frame resized so that the short side is 360 to fit the visual encoder; during training at most 12 frames are kept per video, with at most 1,024 visual tokens per frame
- **Augmented prompt with expected outcome**: identical to the evaluation prompt
- **Structured score supervision**: JSON with only the target key and its 1–5 human score

I EVALUATOR CONFIGURATION DETAILS

This appendix describes the prompt structure of the VLM judge, ablates the schema choice against alternative designs, and ablates the inference-time frame sampling rate.

I.1 SYSTEM PROMPT

You are a strict video evaluation model.

I.2 Scoring Dimensions and Prompt Structure

Each dimension (general or physical law) is a single, independent VLM call, and the VLM directly outputs a JSON object with a 1–5 score. The sub-questions are provided to the VLM as a mental checklist (hint) before scoring, but the VLM is not required to answer each one individually.

I.2.1 General Dimensions

Each general dimension (SA, PTV, Object Persistence; Section 3.4) is prompted with the augmented prompt, the dimension's sub-questions, and the 1–5 scale anchors.

I.3 Schema Ablation

We ablate the prompt schema that elicits per-law scores from the judge: the bare rubric (no sub-questions), per-law sub-questions (+SubQ), per-law sub-questions with chain-of-thought (+SubQ +CoT), and free-form chain-of-thought without sub-questions (+CoT). Table 25 reports the aggregate relative bias of Table 5, here pooled per domain, against human labels per general dimension (SA/PTV/Persist.) and per physics domain (Solid-Body/Fluid/Optical). The judges are Gemini-3.1-Pro and, as open-weights references, Qwen9B-base and Qwen27B-base.

Table 25: Absolute grouped mean-bias $E = |\bar{J} - \bar{H}|/\bar{H}$ (in %) of three zero-shot judges under four prompt schemas, by general dimension and by physics domain; each domain column pools all law–video pairs in that domain. Schemas: the bare rubric, free-form chain-of-thought (+ CoT), per-law sub-questions (+ SubQ), and both. **Bold** marks the column minimum per judge.

Model	SA ↓	PTV ↓	Persist. ↓	Solid-Body ↓	Fluid ↓	Optical ↓
Gemini-3.1-Pro	**7.4**	15.9	39.2	40.2	47.9	27.1
+ CoT	15.1	**0.9**	34.4	31.4	44.2	25.9
+ SubQ	25.4	9.4	**17.3**	22.4	**7.8**	9.4
+ SubQ + CoT	12.6	10.4	28.5	**9.9**	20.6	**1.8**
Qwen9B-base	29.1	44.1	34.0	**47.8**	**22.8**	**20.0**
+ CoT	**23.9**	42.8	40.4	53.8	28.2	27.1
+ SubQ	27.1	**26.8**	**30.9**	55.6	27.0	24.7
+ SubQ + CoT	28.1	55.9	41.1	62.2	30.5	35.3
Qwen27B-base	30.7	39.6	40.1	**48.4**	33.3	**27.6**
+ CoT	**21.0**	33.8	22.8	57.2	**32.7**	32.9
+ SubQ	22.0	**24.6**	**17.1**	56.7	34.8	32.9
+ SubQ + CoT	33.7	49.2	37.1	68.5	37.4	36.5

For Gemini-3.1-Pro, +SubQ lowers the mean-bias on every physics domain, and on Persistence, Fluid and Optical by more than half relative to the bare rubric. On the zero-shot Qwen backbones the same swap lowers bias on the general dimensions but raises it on every physics domain, for both 9B and 27B; Appendix F ablates the schema on the fine-tuned judge. Free-form CoT alone helps Gemini on every physics domain and hurts Qwen on essentially every one. These domain-pooled biases are not directly comparable to the Physics column of Table 5, which macro-averages per-law biases instead of pooling samples.

Stacking chain-of-thought on sub-questions. Stacking chain-of-thought on top is non-monotone: Solid-Body and Optical improve further, but Fluid regresses. We therefore keep +SubQ without CoT as the default schema and treat +SubQ +CoT as a configuration to enable on a per-domain basis.

I.4 Frame Sampling Rate

We ablate the inference frame-sampling rate of PhyJudge-9B at fps $\in \{1, 2, 4, 8\}$, with the zero-shot Qwen9B-base and Qwen27B-base as references. Per-dimension aggregate relative bias against human labels (Table 26) is minimized at 4fps on the General, Overall, and signed columns; on Physics, 1fps and 4fps are close and both well below 2fps and 8fps. The main results use 4fps; relative to 8fps it also halves the visual token input, lowering inference cost and latency.

Table 26: Frame-rate ablation under the sub-question schema. General, Physics and Overall are the absolute grouped mean-bias $E = |\bar{J} - \bar{H}|/\bar{H}$ against human labels, macro-averaged over the general dimensions, over the physics domains, and their equal-weight mean; Signed is the signed relative bias of the per-video overall score. **Bold** marks the best rate per judge and column.

Judge	FPS	**General (%)** ↓	**Physics (%)** ↓	**Overall (%)** ↓	**Signed (%)** → 0
PhyJudge-9B	1	3.0	**4.8**	3.9	-2.8
	2	2.5	8.2	5.4	-1.4
	4	**1.1**	5.5	**3.3**	**-1.1**
	8	3.8	9.1	6.4	-4.9
Qwen9B-base	1	**27.5**	35.5	31.5	+38.1
	2	28.4	**31.4**	**29.9**	**+36.9**
	4	28.2	35.1	31.6	+37.7
	8	30.0	31.8	30.9	+37.8
Qwen27B-base	1	21.5	41.2	31.3	+36.5
	2	23.2	**40.7**	32.0	+36.1
	4	**21.4**	40.8	**31.1**	**+35.7**
	8	24.0	41.0	32.5	+38.0

I.5 Evaluation Cost

Table 27 summarises the approximate API and compute cost of the PhyGround automatic evaluation. The Claude (Opus 4.7) chain-of-thought judge dominates the total because of its image-token pricing.

Table 27: Approximate cost of the automatic evaluation pipeline. *VLM judges* are the cloud judges' per-video scoring calls; *video generation* covers the closed-source generator's outputs.

Item	Cost (USD)
VLM judges	
Gemini-3.1-Pro	∼420
Claude (Opus 4.7)	∼972
Video generation	
Veo-3.1	∼100
Total	**∼1,492**

J Extended Judge Validation

The main text validates PhyJudge-9B by aggregate relative bias on a single held-out generator (Table 5). Aggregate bias measures calibration; this appendix adds per-video error, ranking at the granularity of the paper's conclusions, and transfer to unseen generators, prompts and criteria.

J.1 Same-Backbone Control

The controlled comparison for the fine-tuning claim is the same backbone without fine-tuning, Qwen9B-base, scored with the identical sub-question prompt and frame rate. Fine-tuning cuts aggregate bias by an order of magnitude (Table 5) and improves every per-video metric (Table 28).

J.2 Per-Video Error

Table 28 reports per-video error for all nine judges of Table 5, on the 250 held-out Veo-3.1 videos shared by every judge, against the per-video annotator median (the judge's training and evaluation reference). PhyJudge-9B has the lowest MAE and RMSE and the highest within-± 1 rate.

The reference is a median of 2 to 3 annotator ratings, and a single annotator scored against the median of the video's remaining annotators reaches only Spearman $\times 100$ 28.5, so per-video rank correlation is bounded by reference noise; ranking is validated at the granularities below.

Table 28: Per-video error against the per-video annotator median on the held-out Veo-3.1 videos shared by all judges. Overall is the equal-weight mean of General and Physics per video; within-$\pm t$ is the share of videos with $|\text{judge} - \text{human}| \leq t$. **Bold** marks the best judge per column.

Judge	MAE ↓	RMSE ↓	within-$\pm$0.5 (%)	within-$\pm$1 (%)
Claude (Opus 4.7)	0.82	1.02	**37.6**	67.2
Claude (Opus 4.7, CoT)	1.17	1.40	23.6	48.8
Gemini-3.1-Pro	1.05	1.27	28.0	54.8
Gemini-3.1-Pro (CoT)	0.87	1.08	**37.6**	62.4
Qwen9B-base	1.29	1.53	18.4	44.4
Qwen9B-base (CoT)	1.53	1.75	12.4	31.6
Qwen27B-base	1.19	1.43	23.2	48.0
Qwen27B-base (CoT)	1.52	1.74	15.2	29.6
PhyJudge-9B	**0.79**	**0.94**	34.4	**68.8**

Table 29: Reproduction of the human (generator, law) profile matrix, whose cells are a generator's mean score on one law. Columns are pooled correlations ($\times 100$) of a judge's matrix with the human one, for judges run on every generator. The last column is PhyJudge-9B's Spearman advantage over each baseline, the difference of the two Spearman cells, with its video-level basic bootstrap interval.

Judge	Spearman	Pearson	PhyJudge-9B advantage [95% CI]
PhyJudge-9B	**69.1**	**70.3**	
Claude (Opus 4.7, CoT)	45.9	45.0	+23.2 [17.1, 43.0]
Gemini-3.1-Pro (CoT)	33.2	35.5	+35.9 [31.7, 57.9]

J.3 Ranking at the Granularity of the Paper's Conclusions

The benchmark draws its conclusions from the (generator, law) profile matrix (Table 19 shows it with thin laws merged) and the leaderboard. Table 29 tests whether each judge reproduces the human profile matrix: 8 generators $\times$ 13 laws, each cell the mean over the generator's videos carrying that law, for the three judges with complete 8-generator runs. PhyJudge-9B reproduces the matrix best, and the bootstrap interval of its advantage excludes zero against both baselines.

J.4 Leave-One-Generator-Out

The main-text protocol holds out one generator (Veo-3.1). To remove single-split dependence we retrain eight times, each time excluding one generator from training, and score that generator's 250 videos with the judge that never saw it (*Unseen*) and with a matched judge trained on all eight generators with identical data and hyperparameters (*Seen*). Table 30 reports every fold.

Table 30: Leave-one-generator-out. For each generator, the *Unseen* judge is retrained with that generator excluded and scored on its videos; the *Seen* judge is trained identically but on all eight generators. Per-video Overall: MAE, within-$\pm$1, and signed bias (judge mean $-$ human mean). The mean row averages unrounded fold values; its bias entries are mean absolute signed bias.

	MAE ↓		within-$\pm$1 (%)		Signed bias → 0	
Held-out generator	Unseen	Seen	Unseen	Seen	Unseen	Seen
Veo-3.1	0.78	0.75	68.8	68.0	−0.17	+0.38
OmniWeaving	0.67	0.69	77.2	75.6	+0.05	+0.36
Cosmos-2B	1.07	1.00	50.0	53.6	+0.86	+0.86
Wan2.2-TI2V-5B	0.84	0.93	67.6	56.8	+0.58	+0.75
Wan2.2-27B-A14B	0.75	0.65	71.6	79.2	−0.28	+0.12
Cosmos-14B	0.74	0.85	70.0	64.4	+0.46	+0.66
LTX-2-19B	0.78	0.78	70.4	67.2	+0.30	+0.50
LTX-2.3-22B	0.75	0.76	69.6	69.2	+0.30	+0.41
Mean (8 folds)	**0.80**	**0.80**	**68.2**	**66.8**	**0.37**	**0.50**

- **Error transfers.** Mean per-video MAE is identical for the two judges, and on physics-pooled (video, law) MAE the Unseen judge is equal or better in 7 of 8 folds.
- **Calibration transfers.** Mean absolute signed bias is lower for the Unseen judge than for the Seen judge, and the Unseen judge is equal or better calibrated in 7 of 8 folds: a judge trained on a generator tends to over-score it slightly more than a judge that has not seen it.

A new generator can therefore be scored without retraining the judge.

J.5 Out-of-Distribution Transfer to VideoPhy-2

We score the full VideoPhy-2 test set (Bansal et al., 2025) with the unchanged PhyJudge-9B and with its zero-shot backbone: 3,397 videos from 7 generators, none of which is a model of the PhyGround leaderboard (VideoPhy-2's Wan and Cosmos predate the Wan2.2 and Cosmos-Predict2.5 releases we evaluate), with human physical-commonsense (PC) labels on a 1–5 scale collected under VideoPhy-2's own criteria. Each video's VideoPhy-2 physical rules are mapped to laws of our taxonomy; the 87% of videos with at least one mappable law form the physics rows. The judge's physics score, the mean over the video's mapped laws, is compared with the human PC label. Table 31 reports the pooled metrics and Table 32 the per-generator view; the hard subset is VideoPhy-2's hard split.

Table 31: Out-of-distribution transfer to the VideoPhy-2 test set, whose generators, prompts and human criteria are all external to PhyGround. The unchanged PhyJudge-9B and its zero-shot backbone score every video with our judge settings; physics rows compare the judge's mean applicable-law score with the human physical-commonsense (PC) label on videos whose VideoPhy-2 rules map to a law of our taxonomy. The deduplicated block drops test videos whose caption matches a suite prompt. Generator ranking is Spearman ρ ($\times 100$) over per-generator means. **Bold** marks the better judge.

Set	Metric	PhyJudge-9B	Qwen9B-base
Full mappable set	Generator-ranking Spearman	**64.3**	46.4
	Per-video MAE $\downarrow$	**1.06**	1.41
	Per-video within-± 1 (%)	**66.2**	54.9
	Per-video MAE, hard subset $\downarrow$	**1.02**	1.61
	Per-video within-± 1, hard subset (%)	**69.3**	45.3
Deduplicated by caption	Generator-ranking Spearman	**71.4**	-10.7
	Per-video MAE $\downarrow$	**1.05**	1.33

Table 32: Generator-level view of the VideoPhy-2 transfer on the full mappable set. Human PC and judge scores are means on the rating scale; relative bias is $(\bar{J} - \bar{H})/\bar{H}$.

		PhyJudge-9B		Qwen9B-base	
Generator	Human PC	Score	Rel. bias	Score	Rel. bias
CogVideoX	3.73	3.19	-14.4%	4.64	$+24.4\%$
Wan	3.66	3.03	-17.3%	4.73	$+29.0\%$
Sora	3.66	3.08	-15.9%	4.59	$+25.3\%$
Cosmos	3.55	3.12	-12.2%	4.65	$+30.9\%$
Ray2	3.37	3.14	-6.8%	4.69	$+39.4\%$
HunyuanVideo	3.18	2.99	-5.9%	4.62	$+45.2\%$
VideoCrafter	3.04	2.82	-7.4%	4.41	$+44.9\%$

- **Every physics metric improves** over the backbone, on the hard subset included.
- **Bias is systematic in sign** on every generator (Table 32).
- **All axes unseen at once.** After caption deduplication against our prompt suite, videos, prompts, labeling criteria and generators are unseen, and the separation widens.

J.6 LEAVE-ONE-LAW-OUT

The taxonomy is the axis a fine-tuned judge is most likely to memorise. To test whether PhyJudge-9B applies a criterion it is handed, we partition the laws into four groups that share observables, delete one group from the training data, retrain, and score that group's laws with a judge that never saw their names, criteria or sub-questions. Holding out a whole correlated group matters: removing collision alone while keeping momentum and impenetrability would test a renamed key. The folds are

- **Contact**: collision, momentum, impenetrability;
- **Trajectory**: gravity, inertia;
- **Fluid**: buoyancy, displacement, flow dynamics, fluid continuity, boundary interaction;
- **Optics and material**: reflection, shadow, material.

The control is a judge trained with identical data and hyperparameters but all 13 laws, evaluated on the identical (video, law) samples, together with the zero-shot backbone. Held-out laws are scored on two splits whose videos appear in no fold's training data: *unseen prompt* (175 videos across the 7 training generators, 601 samples), where the law and the prompt are unseen at once, and *unseen generator* (225 Veo-3.1 videos, 790 samples). The backbone's few unparseable outputs are dropped. Table 33 pools the four folds and Table 34 gives every law.

Table 33: Leave-one-law-out, pooled over the four folds (every law held out once). *Unseen-law judge* is the fold's judge trained without the held-out group; *All-13 control* is trained identically but on all laws; *Base* is the zero-shot backbone. Bias is judge mean − human mean; within-±1 and Exact are sample shares in %; Pearson and Spearman are $\times100$. **Bold** marks the better fine-tuned judge.

Scorer	MAE ↓	RMSE ↓	Bias	within-±1	Exact	Pearson	Spearman
Unseen prompt							
Unseen-law judge	**0.85**	**1.11**	$+0.11$	**83.4**	32.6	21.1	22.2
All-13 control	**0.85**	1.12	$+0.07$	82.9	**33.6**	**22.1**	**24.4**
Base	2.35	2.60	$+2.21$	22.0	5.7	-7.2	-2.8
Unseen generator							
Unseen-law judge	**0.89**	**1.16**	-0.36	**81.3**	**31.3**	23.0	21.5
All-13 control	0.92	1.18	-0.45	79.6	30.6	**24.1**	**22.1**
Base	1.90	2.18	$+1.76$	39.9	8.3	-5.6	-4.8

Table 34: Per-law MAE when the law is held out of training (Unseen) versus the All-13 control and the zero-shot backbone, on both held-out splits; the first column gives the fold each law belongs to.

		Unseen prompt			Unseen generator		
Fold	Law	Unseen	All-13	Base	Unseen	All-13	Base
Contact	Collision	0.79	0.83	2.39	0.94	0.92	2.09
	Momentum	0.80	0.80	2.57	0.95	0.91	2.18
	Impenetrability	0.98	0.96	2.27	0.95	0.92	1.85
Trajectory	Gravity	0.89	0.84	2.62	0.88	0.94	1.94
	Inertia	0.71	0.64	2.46	0.94	0.90	1.98
Fluid	Buoyancy	1.36	1.36	1.79	0.94	1.06	1.31
	Displacement	1.07	1.07	1.79	0.62	0.83	1.39
	Flow dynamics	0.79	0.86	2.57	0.80	0.97	1.27
	Fluid continuity	1.05	1.05	2.00	1.08	1.08	1.48
	Boundary interaction	1.21	1.07	2.32	0.82	0.82	1.50
Optics & material	Reflection	0.79	0.71	1.64	1.13	1.20	1.27
	Shadow	0.60	0.60	1.85	0.63	0.78	1.70
	Material	0.71	0.78	2.31	0.80	0.84	2.02

- **Removing a law from training costs essentially nothing.** On the unseen-prompt split the unseen-law judge and the control barely differ on any metric; on the unseen-generator split the unseen-law judge is slightly better on error.

- **Both far exceed the zero-shot backbone** on every metric.
- **No held-out law degrades**: none trails the control by over 0.14 MAE on either split.

The judge therefore applies a criterion it is handed.

J.7 Sensitivity to the Human Reference

Both the leaderboard and the judge treat aggregated human ratings as the reference. We recompute both under nine reference variants, each used in turn as the reference: two aggregations of the full pool (mean, and median; for the judge this is the integer median used for its training, which rounds an even-count tie down), two education subgroups, two major subgroups, and three recruitment-cohort references (the four cohorts of Appendix B, with the two smallest pooled). Subgroup references come from fewer annotators and cover fewer Veo-3.1 videos, so judge bias is computed on the intersection with the judge's predictions. The paper's headline row is the full-pool median, and its small residual against Table 5 comes from per-judge versus all-judge common video sets.

Table 35: Leaderboard sensitivity to the choice of human reference. The baseline is the full post-QC pool with per-(video, dimension) mean aggregation; for each variant, Spearman ρ and Kendall τ of its Overall ranking against the baseline ranking, and the largest per-model shift in Overall score.

Family	Reference variant	Spearman ρ	Kendall τ	max $\lvert\Delta\rvert$
Aggregation	Full pool, mean (baseline)	1.00	1.00	0.00
	Full pool, median	1.00	1.00	0.02
Education	Graduate only (MS, PhD)	1.00	1.00	0.09
	Bachelor only	0.98	0.93	0.04
Major	STEM majors	0.98	0.93	0.18
	Non-STEM majors	0.98	0.93	0.06
Cohort	Cohort 1	1.00	1.00	0.17
	Cohort 2	0.98	0.93	0.09
	Cohort 3	0.98	0.93	0.13

Table 36: Judge aggregate relative bias (%) on held-out Veo-3.1, as General / Physics / Overall, when each human-reference variant is used in turn as the reference.

Family	Reference variant	PhyJudge-9B	Gemini-3.1-Pro
Aggregation	Full pool, mean	5.0 / 7.5 / 6.2	15.8 / 18.0 / 16.9
	Full pool, median	1.1 / 5.5 / 3.3	17.4 / 15.9 / 16.7
Education	Graduate only (MS, PhD)	2.7 / 7.1 / 4.9	18.7 / 17.8 / 18.2
	Bachelor only	6.0 / 7.2 / 6.6	13.8 / 17.1 / 15.4
Major	STEM majors	5.2 / 8.2 / 6.7	16.0 / 15.8 / 15.9
	Non-STEM majors	5.2 / 7.7 / 6.4	15.0 / 18.0 / 16.5
Cohort	Cohort 1	8.9 / 13.8 / 11.4	16.1 / 18.8 / 17.4
	Cohort 2	4.6 / 7.1 / 5.9	13.4 / 18.1 / 15.7
	Cohort 3	4.4 / 10.3 / 7.4	20.6 / 20.5 / 20.5

- **Leaderboard** (Table 35). The 8-model Overall ranking holds under all nine references, and each variant induces at most one adjacent swap. The only movement is between Wan2.2-27B-A14B and Veo-3.1 at the top, which tie on Overall in Table 2.
- **Judge** (Table 36). PhyJudge-9B's Overall bias is lower than Gemini-3.1-Pro's under every reference, including the graduate-only reference closest to an expert panel.

J.8 Use Beyond the Taxonomy

A new phenomenon is added through the schema that produced the 13 laws: add the law, write its 2 to 3 observable sub-questions (Table 10), collect human annotations for it, and re-validate the judge

against them. Appendix J.6 shows that a judge handed an unseen criterion in this schema scores it about as well as one trained on it. The 13 laws cover the three observable domains that carry most physical events in video-generation prompts; growing the taxonomy beyond them is future work.

K Video Generation Inference Settings

Table 37 lists the inference settings of the eight video generation models. Every model receives the same augmented prompt (Appendix A.1), so text conditioning is constant. For Veo-3.1, FPS, clip length, frame count, and resolution are recovered by ffprobing the returned video files.

Table 37: Inference settings of the video models. FPS, length, frames, classifier-free guidance scale (CFG) and sampling steps follow each model's released defaults; the closed Veo-3.1 API exposes none of them (–). For aspect-ratio-adaptive models, the listed resolution is the most common output.

Model	FPS	Length (s)	Resolution	Frames	CFG	Steps	Scheduler
Cosmos-2B	16	5.8	1280x704	93	7	36	FlowMatchEulerDiscrete
Cosmos-14B	16	5.8	1280x704	93	7	36	FlowMatchEulerDiscrete
LTX-2-19B	25	3.9	832x448	97	3	30	FlowMatchEulerDiscrete
LTX-2.3-22B	25	3.9	832x448	97	3	30	FlowMatchEulerDiscrete
OmniWeaving	16	5.1	848x480	81	6	50	FlowMatchEulerDiscrete
Wan2.2-TI2V-5B	16	5.1	1248x704	81	5.5	50	FlowUniPCMultistep
Wan2.2-27B-A14B	16	5.1	832x464	81	3.5	40	FlowUniPCMultistep
Veo-3.1	24	8.0	1280x720	192	–	–	–

The two Cosmos models and Veo-3.1 write a fixed output resolution regardless of input, while the remaining models resize to track the input aspect ratio. All open models run in `bf16`. For the LTX-2 family we run only the base denoising pass and skip the released distilled-LoRA refinement pass.

L Scope of the Measurement

Outcome-conditioned physical adherence. Every prompt states the expected outcome (Section 3.1). PhyGround thus measures whether a model can plausibly realize a specified event. This makes per-law evaluation well-posed: annotator and judge verify an observable consequence.

For the prompt "a glass is knocked off a table", suppose the generated video shows the glass landing intact. With the stated outcome "the glass shatters on the floor", this is unambiguously a failure to realize the specified event. Without it, one annotator reads the same clip as a rigid-body violation while another reads it as a valid outcome for a sturdy glass. That ambiguity lowers inter-annotator agreement, injects noise into the per-law labels the judge is trained on, and turns the 1–5 severity scale into a mixture of physics judgment and outcome guessing.

Explicit outcomes remove ambiguity from unobserved material and impact conditions while still requiring coherent dynamics. A model may realize "a glass falls and shatters" while making the glass hover or shatter before contact; it follows the instruction and remains physically implausible. Semantic adherence is therefore necessary but not sufficient for a high PhyGround score.

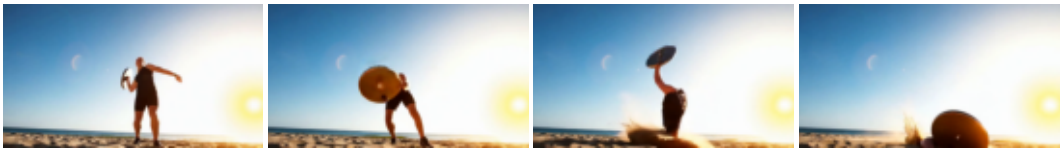
Structured perceived adherence. PhyGround is a structured human-perception benchmark: scores are judgments of law-specific observables. Two mechanisms discipline visual judgment toward those observables (fall direction, liquid continuity, shadow and reflection consistency): annotators rate against each law's anchor question with concrete observation notes, and the VLM judge is prompted with each law's 2 to 3 observable sub-questions. Human ratings are the reference for observable, law-specific adherence, the same reference that VideoPhy-2 (Bansal et al., 2025) and WorldModelBench (Li et al., 2025) adopt, and Appendix J.7 shows the measurement is stable across subgroups and aggregations of the annotator pool.

Instrumented measurement would not deliver correctness by itself. Generated videos provide no camera calibration or ground-truth state to measure against, so trajectories, optical geometry and fluid state are underdetermined at the source; and even where one quantity is measurable, its consistency is

neither necessary nor sufficient for physical correctness. A clip can keep shadow geometry or liquid volume internally consistent while the depicted event is physically wrong. In the other direction, frame-consistency metrics do detect appearance change, but they score liquid appearing out of nowhere (Figure 29) the same as legitimate physical change such as a glass shattering or a liquid splashing. Separating violation from valid dynamics requires judging the event itself.

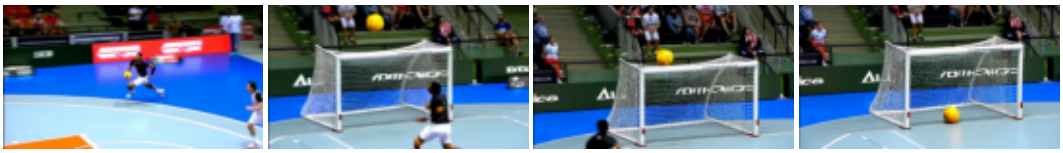
M Qualitative Examples of Physical Law Violations

Figures 19 to 31 show qualitative examples of violations of each physical law. Every video shown received a score of 1 on the labeled law from every annotator who rated it.



(a) Model: Cosmos-2B

Prompt: A discus is thrown with an overhand motion, traveling a significant distance before landing in the soft sand.



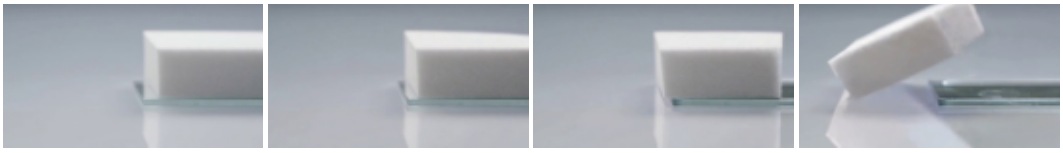
(b) Model: Cosmos-2B

Prompt: A player runs and hits the ball into the goal, the ball directly impacting the backboard.



(c) Model: Cosmos-2B

Prompt: A person vigorously twists a wet towel, water spraying outwards in a visible arc.



(d) Model: Cosmos-14B

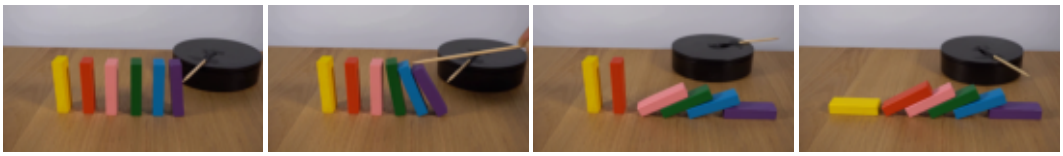
Prompt: A small, square piece of foam is placed on a slanted acrylic surface; it slides down, creating a soft impact sound upon landing.



(e) Model: Cosmos-2B

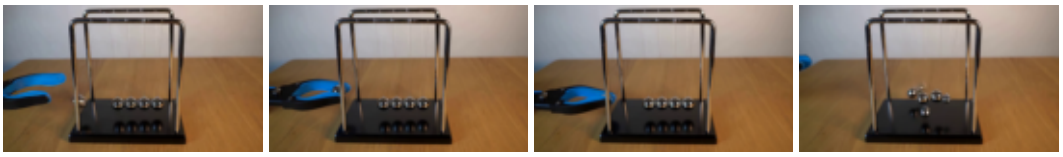
Prompt: A long jumper demonstrates a flawed technique, resulting in a short distance and an unbalanced landing.

Figure 19: **Gravity** violations: unsupported objects hover or accelerate unlike free fall.



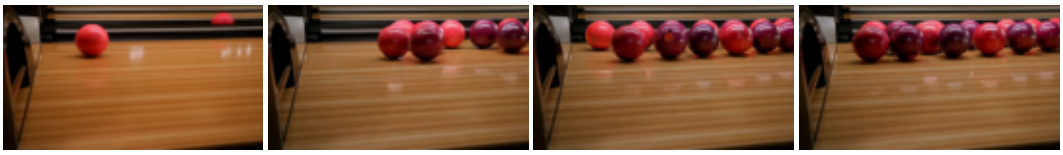
(a) Model: Veo-3.1

Prompt: A row of colorful wooden blocks lined up on a wooden table with a wooden stick attached to a black rotating platform. The platform rotates clockwise and the wooden stick hits the first block as it rotates. The blocks topple over one by one in a domino chain reaction. Static shot with no camera movement.



(b) Model: OmniWeaving

Prompt: A Newton's cradle device on the table and one of the metal balls is held up by a blue handled grabber tool. The claw releases the ball. The released ball swings down, hits the stationary balls, and the ball on the opposite end swings out. Static shot with no camera movement.



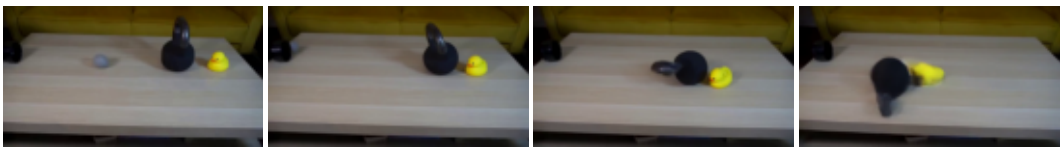
(c) Model: Cosmos-14B

Prompt: Multiple bowling balls are rolled consecutively, leaving visible marks on the lane.



(d) Model: LTX-2-19B

Prompt: An olive green fabric box is on a light wood surface. A brown tennis ball rolls out of the black tube sitting on the table and rolls towards the box. The tennis ball impacts the fabric box, pushing it across the surface. Static shot with no camera movement.



(e) Model: LTX-2-19B

Prompt: A light beige coffee table with a black kettlebell and a yellow rubber duck on it. A grey tennis ball rolls out of the black tube sitting on the table and towards the duck and kettlebell. The tennis ball collides with the rubber duck, pushing it across the table. Static shot with no camera movement.

Figure 20: **Inertia** violations: objects start or stop moving without an external force.



(a) Model: LTX-2-19B

Prompt: A player hits the ball off the side of the mallet, resulting in an unexpected trajectory.



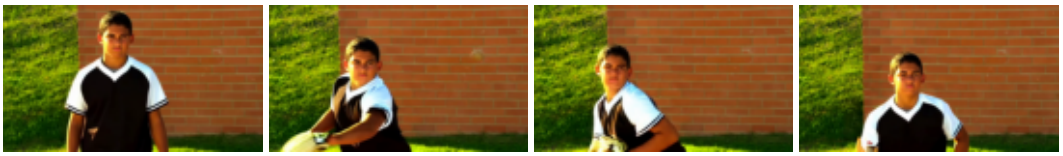
(b) Model: Cosmos-2B

Prompt: A person drop kicks a pumpkin across a cobblestone street, the pumpkin rolling before stopping.



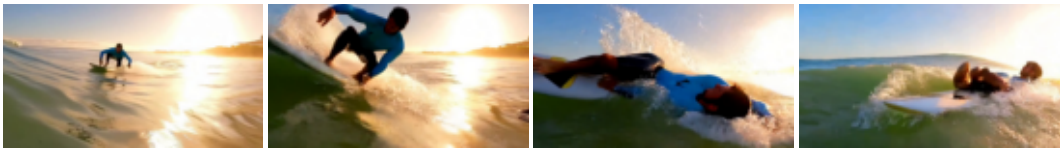
(c) Model: LTX-2-19B

Prompt: A player kicks a kickball that bounces off a wooden fence, changing direction.



(d) Model: Wan2.2-TI2V-5B

Prompt: A player throws a softball against a brick wall; the ball rebounds at a visible angle.



(e) Model: Cosmos-2B

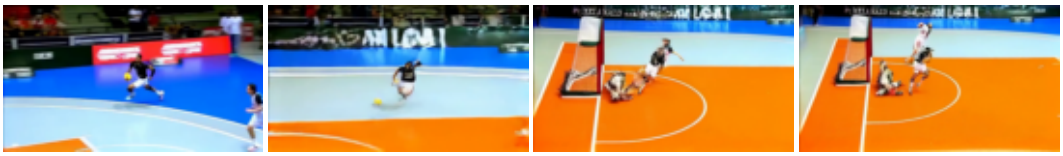
Prompt: A surfer falls off their board, colliding briefly with another surfer, before swimming away.

Figure 21: **Momentum** violations: post-collision motion contradicts incoming momentum.



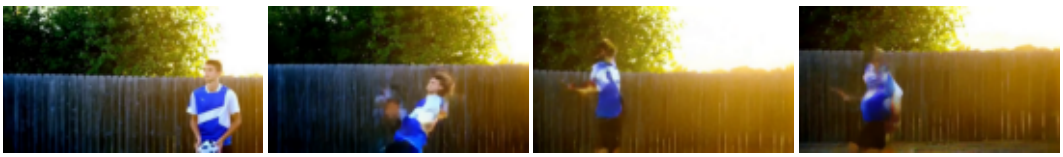
(a) Model: LTX-2-19B

Prompt: A player hits the ball off the side of the mallet, resulting in an unexpected trajectory.



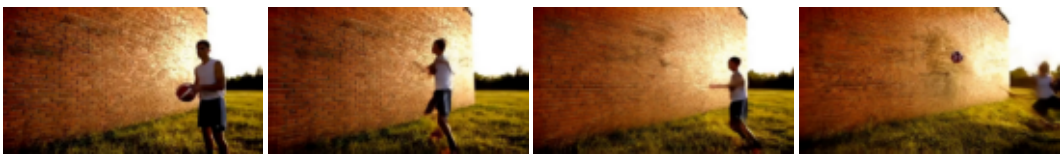
(b) Model: LTX-2-19B

Prompt: A player runs and hits the ball into the goal, the ball directly impacting the backboard.



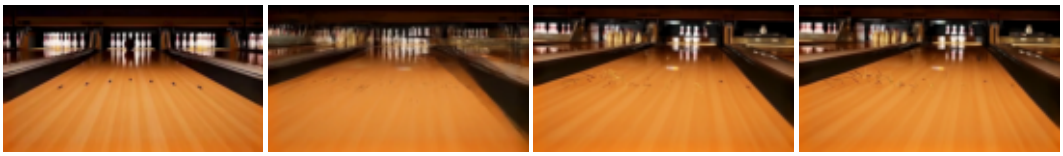
(c) Model: LTX-2-19B

Prompt: A player kicks a kickball that bounces off a wooden fence, changing direction.



(d) Model: LTX-2-19B

Prompt: A kickball is kicked with force, striking a brick wall and leaving a noticeable scuff mark.



(e) Model: LTX-2-19B

Prompt: A bowling ball rolls down a polished wooden lane, hitting the pins at the end.

Figure 22: **Impenetrability** violations: solid objects pass through or overlap one another.



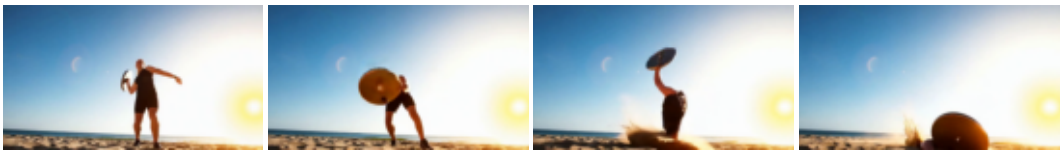
(a) Model: LTX-2-19B

Prompt: A player hits the ball off the side of the mallet, resulting in an unexpected trajectory.



(b) Model: Cosmos-2B

Prompt: A person drop kicks a pumpkin across a cobblestone street, the pumpkin rolling before stopping.



(c) Model: Cosmos-2B

Prompt: A discus is thrown with an overhand motion, traveling a significant distance before landing in the soft sand.



(d) Model: Cosmos-2B

Prompt: A batsman hits a six, the ball clearing the boundary rope.



(e) Model: Cosmos-2B

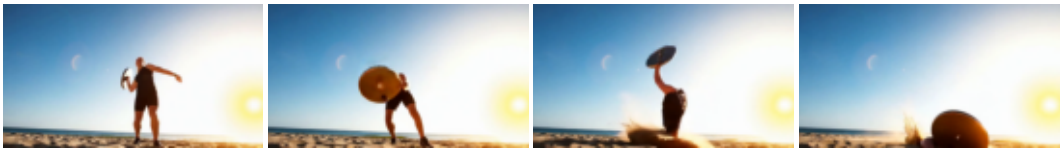
Prompt: A discus hits a large wooden target with a clear impact mark.

Figure 23: **Collision** violations: objects fail to rebound, fracture, or deform properly on impact.



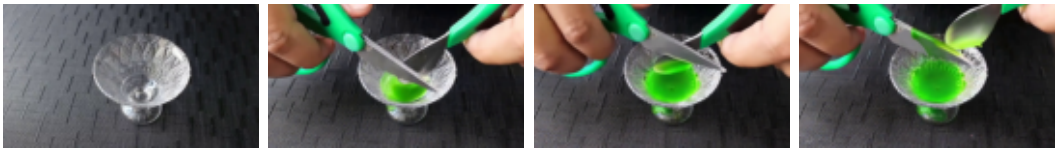
(a) Model: Cosmos-2B

Prompt: A person drop kicks a pumpkin across a cobblestone street, the pumpkin rolling before stopping.



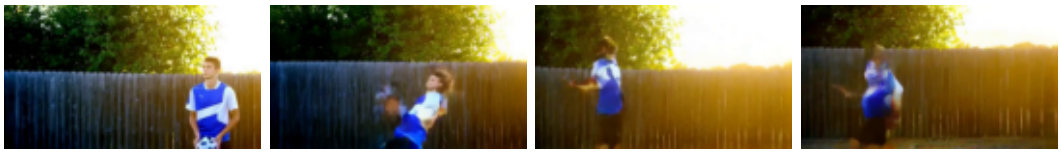
(b) Model: Cosmos-2B

Prompt: A discus is thrown with an overhand motion, traveling a significant distance before landing in the soft sand.



(c) Model: Wan2.2-TI2V-5B

Prompt: A green liquid is poured steadily downward in a continuous stream, pooling inside a clear glass bowl on a black surface.



(d) Model: LTX-2-19B

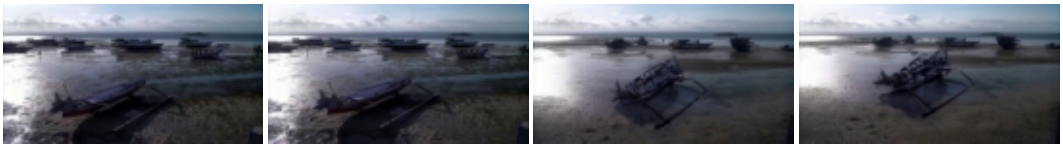
Prompt: A player kicks a kickball that bounces off a wooden fence, changing direction.



(e) Model: Cosmos-2B

Prompt: A person vigorously twists a wet towel, water spraying outwards in a visible arc.

Figure 24: **Material** violations: material-specific responses do not match physical properties (e.g., glass that does not shatter, rubber that does not bounce).



(a) Model: LTX-2-19B

Prompt: The rising ocean tide floods a sandy beach, surrounding several grounded boats. As the water deepens, it lifts the vessels off the wet sand, causing them to become buoyant and float. The boats bob gently on the surface of the incoming water while a person walks along the shoreline.



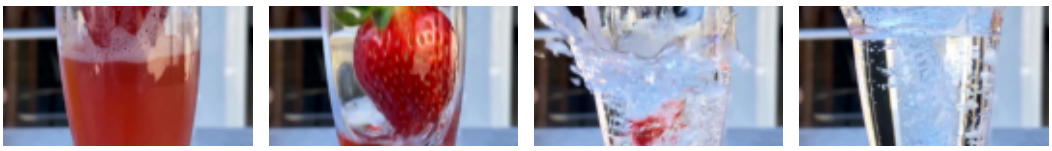
(b) Model: Cosmos-2B

Prompt: A pair of black leather boots is slowly lowered into a clear glass bowl filled with water. As the boots submerge, they displace the liquid, causing the water level to steadily rise until it reaches the top rim of the bowl.



(c) Model: Cosmos-2B

Prompt: A bundle of lit matchsticks is placed in a bowl of red liquid. A glass jar gets lowered and covers the matchsticks. The burning matches consume oxygen, creating a vacuum that pulls the liquid into the jar. Static shot with no camera movement.



(d) Model: Wan2.2-TI2V-5B

Prompt: A red strawberry drops into a glass filled with foamy pink liquid. As the fruit submerges, it displaces the drink, forcing the liquid level to rise and the surface foam to splash outward.



(e) Model: LTX-2-19B

Prompt: A potato is held by a grabber tool and dropped into a tall glass containing blue liquid with a band of green tape marking a level on the glass. The potato displaces the liquid, causing the liquid level to rise above the tape mark. Static shot with no camera movement.

Figure 25: **Buoyancy** violations: dense objects fail to sink or lightweight objects fail to float.



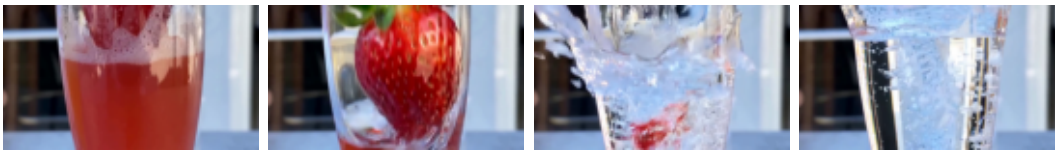
(a) Model: Cosmos-2B

Prompt: A bundle of lit matchsticks is placed in a bowl of red liquid. A glass jar gets lowered and covers the matchsticks. The burning matches consume oxygen, creating a vacuum that pulls the liquid into the jar. Static shot with no camera movement.



(b) Model: Cosmos-2B

Prompt: A glass bottle filled with dark amber syrup sits on a rustic wooden table. In the foreground, a wooden spoon containing the same thick liquid tilts downward. The viscous syrup flows smoothly off the edge of the spoon in a slow, continuous stream.



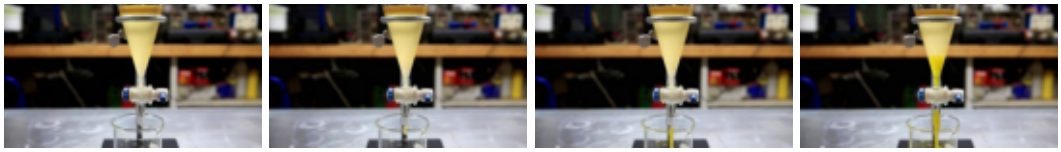
(c) Model: Wan2.2-TI2V-5B

Prompt: A red strawberry drops into a glass filled with foamy pink liquid. As the fruit submerges, it displaces the drink, forcing the liquid level to rise and the surface foam to splash outward.



(d) Model: LTX-2.3-22B

Prompt: A potato is held by a grabber tool and dropped into a tall glass containing blue liquid with a band of green tape marking a level on the glass. The potato displaces the liquid, causing the liquid level to rise above the tape mark. Static shot with no camera movement.



(e) Model: Wan2.2-27B-A14B

Prompt: A glass separatory funnel holds a layered yellow and amber liquid directly above an empty glass beaker. The blue-knobbed valve at the base of the funnel opens, releasing a steady stream of the pale yellow fluid. The liquid flows smoothly downward, splashing into and slowly filling the beaker below.

Figure 26: **Displacement** violations: liquid level fails to rise or full containers do not overflow.



(a) Model: Wan2.2-TI2V-5B

Prompt: A chef in a modern kitchen tips a white bowl, pouring its contents in a steady stream that flows smoothly down into a pot on the stove. He then stirs the mixture with a spoon, causing the liquid to swirl and flow in continuous, circular currents.



(b) Model: Wan2.2-TI2V-5B

Prompt: A heavy car tire rolls over a green bottle on concrete, crushing it flat. The trapped liquid bursts from the broken bottle and rapidly flows outward, spreading across the ground in an expanding puddle.



(c) Model: Cosmos-2B

Prompt: The rising ocean tide floods a sandy beach, surrounding several grounded boats. As the water deepens, it lifts the vessels off the wet sand, causing them to become buoyant and float. The boats bob gently on the surface of the incoming water while a person walks along the shoreline.



(d) Model: Cosmos-14B

Prompt: A glass bottle tips, pouring a thick, golden liquid into a white mug resting on a wooden cutting board. The syrupy liquid flows continuously in a smooth, heavy stream, slowly pooling at the bottom of the mug.



(e) Model: Cosmos-2B

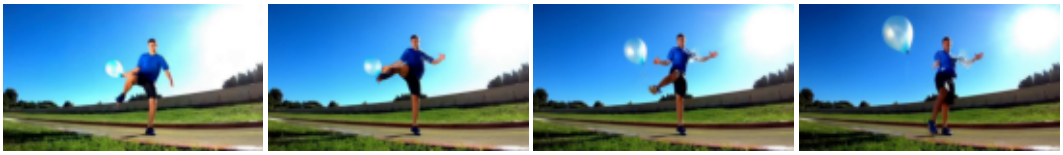
Prompt: A bundle of lit matchsticks is placed in a bowl of red liquid. A glass jar gets lowered and covers the matchsticks. The burning matches consume oxygen, creating a vacuum that pulls the liquid into the jar. Static shot with no camera movement.

Figure 27: **Flow dynamics** violations: liquid spreads, converges, or drains unrealistically.



(a) Model: OmniWeaving

Prompt: A person mops up a puddle of water on a concrete floor, the water disappearing as the mop passes.



(b) Model: LTX-2.3-22B

Prompt: A drop kick sends a water balloon arcing through the air; the balloon bursts upon impact with the ground.



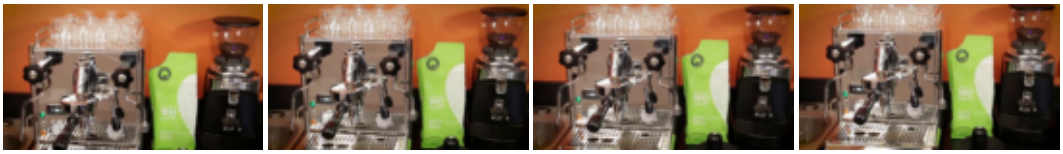
(c) Model: Cosmos-2B

Prompt: The rising ocean tide floods a sandy beach, surrounding several grounded boats. As the water deepens, it lifts the vessels off the wet sand, causing them to become buoyant and float. The boats bob gently on the surface of the incoming water while a person walks along the shoreline.



(d) Model: Cosmos-2B

Prompt: A pair of black leather boots is slowly lowered into a clear glass bowl filled with water. As the boots submerge, they displace the liquid, causing the water level to steadily rise until it reaches the top rim of the bowl.



(e) Model: Wan2.2-27B-A14B

Prompt: A shiny silver espresso machine sits against an orange wall, next to a bright green coffee bag and a bean-filled grinder. Dark, rich espresso begins to flow smoothly from the machine's metal spout, streaming downward in a steady, continuous liquid pour.

Figure 28: **Boundary interaction** violations: liquid splashes, deflects, or splits unrealistically.



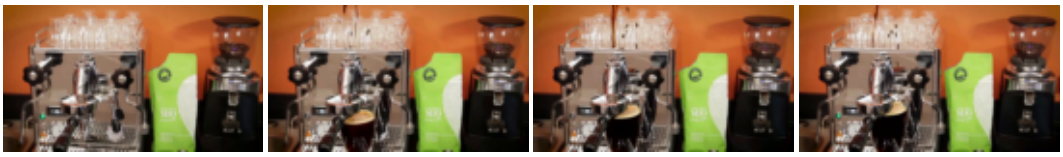
(a) Model: Wan2.2-TI2V-5B

Prompt: A chef in a modern kitchen tips a white bowl, pouring its contents in a steady stream that flows smoothly down into a pot on the stove. He then stirs the mixture with a spoon, causing the liquid to swirl and flow in continuous, circular currents.



(b) Model: Cosmos-14B

Prompt: A heavy car tire rolls over a green bottle on concrete, crushing it flat. The trapped liquid bursts from the broken bottle and rapidly flows outward, spreading across the ground in an expanding puddle.



(c) Model: Wan2.2-TI2V-5B

Prompt: A shiny silver espresso machine sits against an orange wall, next to a bright green coffee bag and a bean-filled grinder. Dark, rich espresso begins to flow smoothly from the machine's metal spout, streaming downward in a steady, continuous liquid pour.



(d) Model: Cosmos-2B

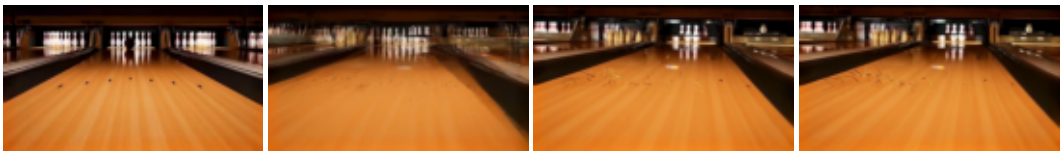
Prompt: A glass bottle filled with dark amber syrup sits on a rustic wooden table. In the foreground, a wooden spoon containing the same thick liquid tilts downward. The viscous syrup flows smoothly off the edge of the spoon in a slow, continuous stream.



(e) Model: Wan2.2-TI2V-5B

Prompt: A woman in a kitchen tilts a bowl, pouring a smooth, steady stream of yellow liquid into a pot below. The thick liquid flows continuously from the rim, pooling inside the cookware. She then drags a wooden spoon through the liquid, stirring it into a swirling motion.

Figure 29: **Fluid continuity** violations: liquid loses mass, breaks up, or appears from nowhere.



(a) Model: LTX-2-19B

Prompt: A bowling ball rolls down a polished wooden lane, hitting the pins at the end.



(b) Model: Cosmos-2B

Prompt: A tennis ball with a smiley face drawn on it is slowly rotating on a black rotating platform that rotates clockwise in front of a mirror and reflecting the side of the ball that has no smiley face. As the ball rotates, its reflection in the mirror will continuously change, eventually revealing the smiley face. Static shot with no camera movement.

Figure 30: **Reflection** violations: reflections do not approximately match the reflected scene.

N Qualitative Examples of Physical Law Violations by Model

Figures 32 to 39 show further examples, grouped by video generation model. Every video shown received a score of 1 on the labeled law from every annotator who rated it.



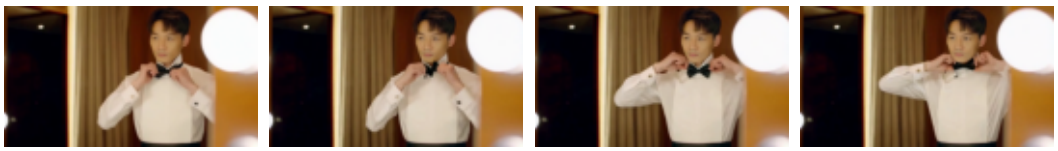
(a) Model: LTX-2-19B

Prompt: The person in blue launches the yellow frisbee across the lawn. The frisbee flies through the air, following a parabolic trajectory before landing on the grass.



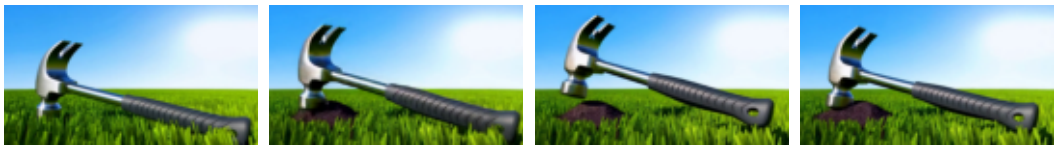
(b) Model: Cosmos-2B

Prompt: A tennis ball with a smiley face drawn on it is slowly rotating on a black rotating platform that rotates clockwise in front of a mirror and reflecting the side of the ball that has no smiley face. As the ball rotates, its reflection in the mirror will continuously change, eventually revealing the smiley face. Static shot with no camera movement.



(c) Model: Cosmos-14B

Prompt: A person ties a bow tie while standing in front of a mirror, adjusting the knot's symmetry.



(d) Model: Cosmos-2B

Prompt: A hammer is thrown, landing in a grassy field, leaving a small crater in the turf.



(e) Model: Wan2.2-27B-A14B

Prompt: A sleek, gold car drives forward along a dusty dirt road under a bright, clear sky. The strong sunlight casts a sharp, dark shadow of the moving vehicle, which shifts and travels across the unpaved ground beside it as the car advances.

Figure 31: **Shadow** violations: shadows contradict the light source or fail to track their object.



(a) Law: Momentum

Prompt: A player hits the ball off the side of the mallet, resulting in an unexpected trajectory.



(b) Law: Boundary interaction

Prompt: A red strawberry drops into a glass filled with foamy pink liquid. As the fruit submerges, it displaces the drink, forcing the liquid level to rise and the surface foam to splash outward.



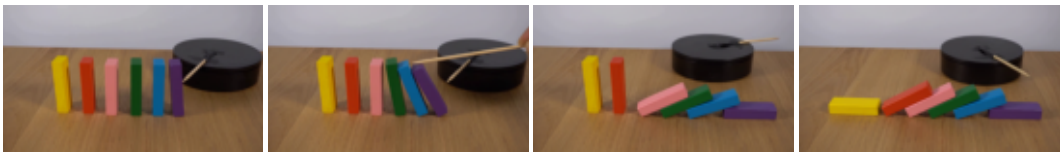
(c) Law: Collision

Prompt: A yellow mug is held by a grabber tool in front of a white projection screen with a concrete brick positioned beneath it. The grabber releases the mug. The mug falls and shatters upon impact with the concrete brick. Static shot with no camera movement.



(d) Law: Impenetrability

Prompt: A heavy bowling pin is nudged with the tip of a bowling ball, causing it to fall.



(e) Law: Inertia

Prompt: A row of colorful wooden blocks lined up on a wooden table with a wooden stick attached to a black rotating platform. The platform rotates clockwise and the wooden stick hits the first block as it rotates. The blocks topple over one by one in a domino chain reaction. Static shot with no camera movement.

Figure 32: Qualitative examples of physical law violations from **Veo-3.1**.



(a) Law: Boundary interaction

Prompt: A person mops up a puddle of water on a concrete floor, the water disappearing as the mop passes.



(b) Law: Collision

Prompt: Two skaters collide, their skates tangling momentarily before they separate.



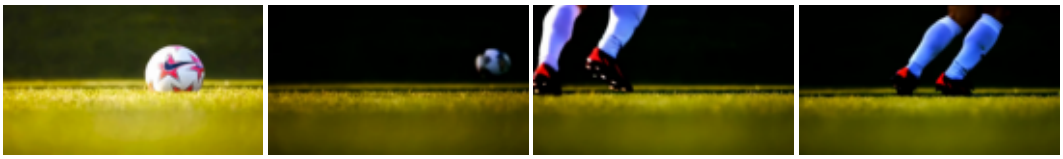
(c) Law: Gravity

Prompt: A player runs and hits the ball into the goal, the ball directly impacting the backboard.



(d) Law: Inertia

Prompt: A Newton's cradle device on the table and one of the metal balls is held up by a blue handled grabber tool. The claw releases the ball. The released ball swings down, hits the stationary balls, and the ball on the opposite end swings out. Static shot with no camera movement.



(e) Law: Momentum

Prompt: A soccer ball is kicked and hits a person's leg, deflecting at a different angle.

Figure 33: Qualitative examples of physical law violations from **OmniWeaving**.



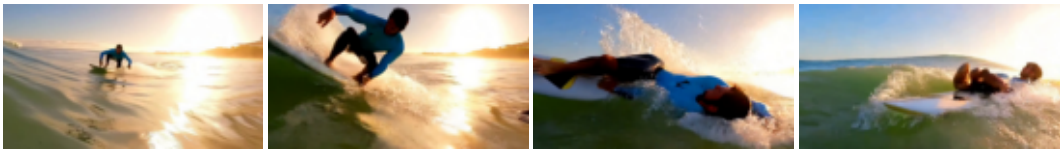
(a) Law: Collision

Prompt: A person drop kicks a pumpkin across a cobblestone street, the pumpkin rolling before stopping.



(b) Law: Gravity

Prompt: A player runs and hits the ball into the goal, the ball directly impacting the backboard.



(c) Law: Momentum

Prompt: A surfer falls off their board, colliding briefly with another surfer, before swimming away.



(d) Law: Boundary interaction

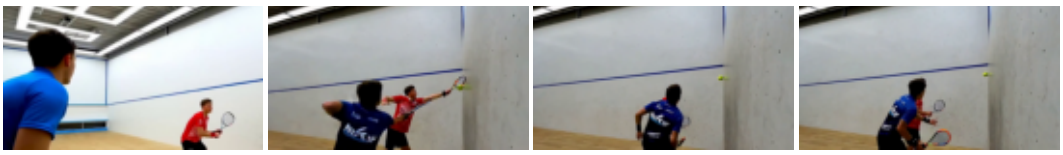
Prompt: The rising ocean tide floods a sandy beach, surrounding several grounded boats. As the water deepens, it lifts the vessels off the wet sand, causing them to become buoyant and float. The boats bob gently on the surface of the incoming water while a person walks along the shoreline.



(e) Law: Buoyancy

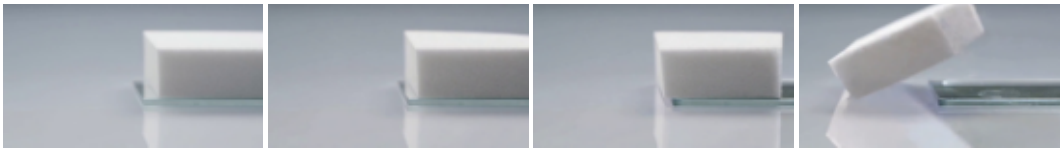
Prompt: A bundle of lit matchsticks is placed in a bowl of red liquid. A glass jar gets lowered and covers the matchsticks. The burning matches consume oxygen, creating a vacuum that pulls the liquid into the jar. Static shot with no camera movement.

Figure 34: Qualitative examples of physical law violations from **Cosmos-2B**.



(a) Law: Collision

Prompt: A player performs a nick shot in squash, the ball barely grazing the side wall before reaching the opponent.



(b) Law: Gravity

Prompt: A small, square piece of foam is placed on a slanted acrylic surface; it slides down, creating a soft impact sound upon landing.



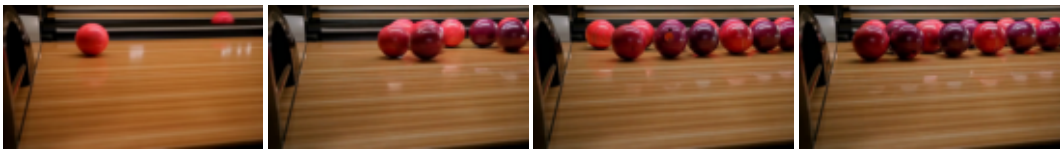
(c) Law: Flow dynamics

Prompt: A glass bottle tips, pouring a thick, golden liquid into a white mug resting on a wooden cutting board. The syrupy liquid flows continuously in a smooth, heavy stream, slowly pooling at the bottom of the mug.



(d) Law: Fluid continuity

Prompt: A heavy car tire rolls over a green bottle on concrete, crushing it flat. The trapped liquid bursts from the broken bottle and rapidly flows outward, spreading across the ground in an expanding puddle.



(e) Law: Impenetrability

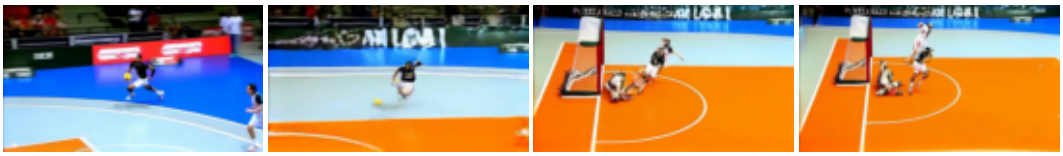
Prompt: Multiple bowling balls are rolled consecutively, leaving visible marks on the lane.

Figure 35: Qualitative examples of physical law violations from **Cosmos-14B**.



(a) Law: Collision

Prompt: A player hits the ball off the side of the mallet, resulting in an unexpected trajectory.



(b) Law: Impenetrability

Prompt: A player runs and hits the ball into the goal, the ball directly impacting the backboard.



(c) Law: Buoyancy

Prompt: The rising ocean tide floods a sandy beach, surrounding several grounded boats. As the water deepens, it lifts the vessels off the wet sand, causing them to become buoyant and float. The boats bob gently on the surface of the incoming water while a person walks along the shoreline.



(d) Law: Flow dynamics

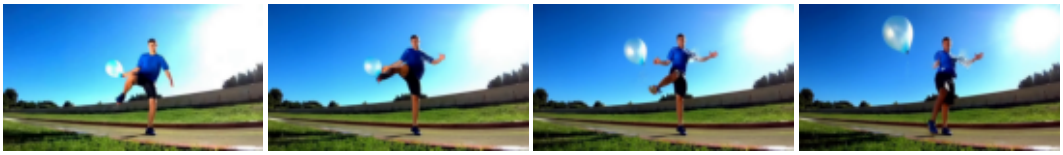
Prompt: Snow melts and slides down the mountain cliff in the scenic landscape.



(e) Law: Gravity

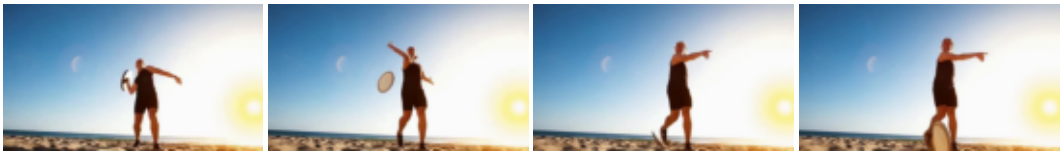
Prompt: A tennis player serves, the ball leaves the racquet at high speed, and lands in the service box.

Figure 36: Qualitative examples of physical law violations from **LTX-2-19B**.



(a) Law: Boundary interaction

Prompt: A drop kick sends a water balloon arcing through the air; the balloon bursts upon impact with the ground.



(b) Law: Collision

Prompt: A discus is thrown with an overhand motion, traveling a significant distance before landing in the soft sand.



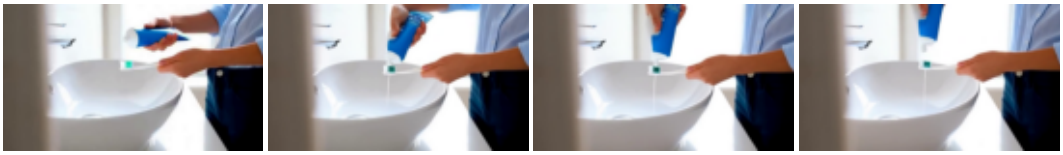
(c) Law: Momentum

Prompt: A player kicks a soccer ball which hits a lamp post, the ball rebounding at an angle.



(d) Law: Displacement

Prompt: A potato is held by a grabber tool and dropped into a tall glass containing blue liquid with a band of green tape marking a level on the glass. The potato displaces the liquid, causing the liquid level to rise above the tape mark. Static shot with no camera movement.



(e) Law: Flow dynamics

Prompt: A person squeezes a tube of toothpaste, a continuous stream emerging onto a toothbrush.

Figure 37: Qualitative examples of physical law violations from **LTX-2.3-22B**.



(a) Law: Gravity

Prompt: An axe is used to fell a small tree, the tree trunk falling to the ground.



(b) Law: Boundary interaction

Prompt: A drop kick sends a water balloon arcing through the air; the balloon bursts upon impact with the ground.



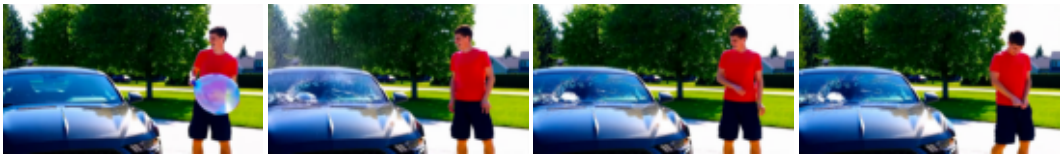
(c) Law: Collision

Prompt: A discus hits a large wooden target with a clear impact mark.



(d) Law: Impenetrability

Prompt: A player runs and hits the ball into the goal, the ball directly impacting the backboard.



(e) Law: Material

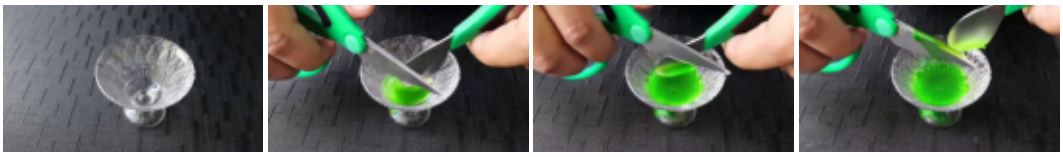
Prompt: Someone throws a water balloon at a car windshield; the balloon bursts, leaving water residue.

Figure 38: Qualitative examples of physical law violations from **Wan2.2-27B-A14B**.



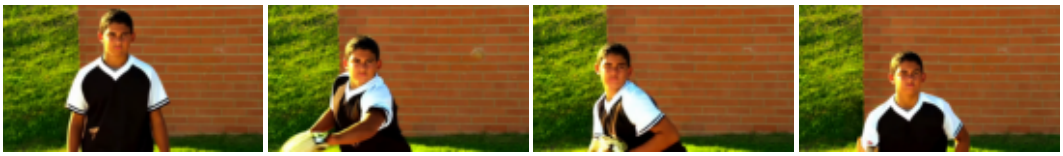
(a) Law: Collision

Prompt: A person throws a frisbee that tumbles along the ground before stopping.



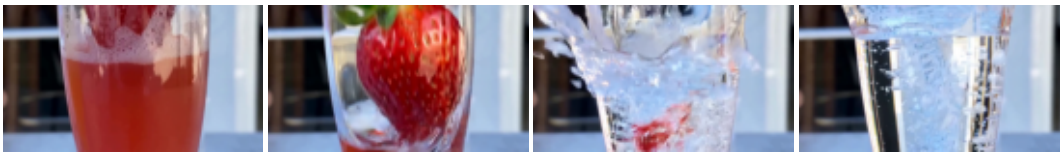
(b) Law: Material

Prompt: A green liquid is poured steadily downward in a continuous stream, pooling inside a clear glass bowl on a black surface.



(c) Law: Momentum

Prompt: A player throws a softball against a brick wall; the ball rebounds at a visible angle.



(d) Law: Buoyancy

Prompt: A red strawberry drops into a glass filled with foamy pink liquid. As the fruit submerges, it displaces the drink, forcing the liquid level to rise and the surface foam to splash outward.



(e) Law: Flow dynamics

Prompt: A chef in a modern kitchen tips a white bowl, pouring its contents in a steady stream that flows smoothly down into a pot on the stove. He then stirs the mixture with a spoon, causing the liquid to swirl and flow in continuous, circular currents.

Figure 39: Qualitative examples of physical law violations from **Wan2.2-TI2V-5B**.