DOES ONE PROMPT REPRESENT A PHYSICAL SCENARIO? PHYSICS EVALUATION THROUGH PROMPT SETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating whether video generation models understand the physical world is essential, since physical understanding is a prerequisite for their use as world models. Existing physics benchmarks generate a video from a text prompt describing a physical scenario and score its physical plausibility. However, each scenario is represented by a single prompt, although it can be described in many semantically and physically equivalent ways. We show that this single-prompt design makes the evaluation sensitive to the choice of prompt. Using a large language model, we rewrite benchmark prompts while preserving their semantics and physics, and evaluate five video generation models on VideoPhy-2 and PhyGenBench with fixed seeds. Rewriting alone changes the physical score in 35.5% and 35.7% of cases, flips the official pass decision in 13.3%, and changes the point ranking of 4 of the 5 models on each benchmark. To address this prompt sensitivity, we propose Prompt Set Evaluation, which scores each scenario over a set of validated equivalent prompts and reports the mean score with its variation; five prompts per scenario substantially improve the stability of scores and rankings. The sensitivity is not random noise, since it persists under fully repeatable evaluators. In a blinded human study of video pairs whose scores differ, raters seldom share the evaluator's preference and judge about half of the pairs equally plausible, which suggests that the vision-language models serving as evaluators also respond to differences between videos unrelated to physics. Physical understanding should therefore be evaluated over a distribution of descriptions rather than a single prompt.

1 INTRODUCTION

Video generation models (Ho et al., 2022b; Singer et al., 2023; Ho et al., 2022a; Blattmann et al., 2023b; Polyak et al., 2024) are increasingly viewed as a path toward world models that simulate how the physical world evolves (Brooks et al., 2024; Bruce et al., 2024; Yang et al., 2024a;b; NVIDIA et al., 2025). For this role, physical understanding is a prerequisite: generated videos must obey the physical constraints of the depicted event, not only look realistic (Lin et al., 2025). Physics benchmarks (Bansal et al., 2025; 2026; Meng et al., 2025; Gu et al., 2026; Guo et al., 2025; Li et al., 2025b; Zheng et al., 2025) therefore describe a physical scenario with a text prompt, generate a video from it, and score its physical plausibility with human raters or, more commonly, an automatic evaluator aligned with human judgment; the scores compare models and measure progress.

However, these benchmarks represent each scenario by a single prompt. A physical scenario can be described in many ways that preserve its objects, actions, and physical conditions; "a leaf blower is pointed at a patch of leaves" and "a leaf blower is aimed at a patch of leaves" describe the same event. Since such descriptions specify the same physics, a model that understands the scenario should generate equally plausible videos for all of them and an evaluator should score them consistently, so the physical score should not depend on which description the benchmark uses. Whether this holds has not been examined, and Figure 1 suggests that it does not: rewriting a prompt changes the physical score and the pass decision while the evaluator's question stays the same.

We show that this single-prompt design makes physics evaluation unreliable. We take 15 scenarios from each of VideoPhy-2 (Bansal et al., 2026) and PhyGenBench (Meng et al., 2025) and use a large

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.

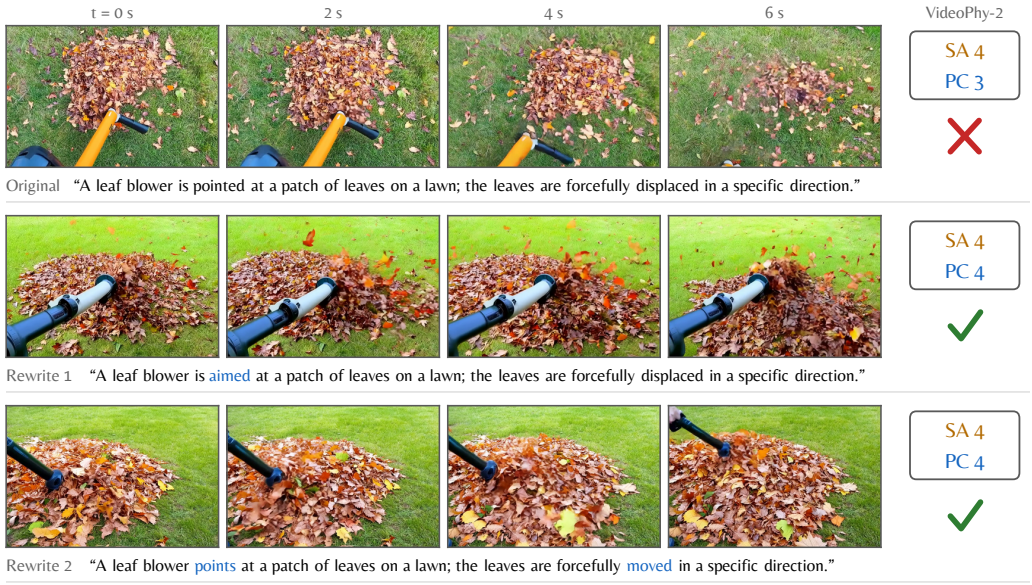


Figure 1: Prompt sensitivity of physics evaluation on VideoPhy-2 (CogVideoX-5B, fixed generation seed). Three descriptions of the same event differ only in the highlighted words. The evaluator asks the same physical-commonsense question of every video and rates semantic adherence (SA) identically, yet the physical-commonsense (PC) score and the pass decision change with the rewrite.

language model to rewrite each prompt five times while preserving its semantics and physics; human annotators and independent language-model judges confirm that the rewrites are equivalent. We then generate videos from the original and rewritten prompts with five video generation models (Yang et al., 2025; Blattmann et al., 2023a; NVIDIA et al., 2025; Wan Team et al., 2025; Kong et al., 2024), using a fixed generation seed, and score them with each benchmark's evaluator as released, up to one Stage-2 prompt variant that does not change the conclusions (Appendix B). The evaluators judge physics without seeing the rewritten prompt, so any change in the physical plausibility score, which we call the physical score, reflects a change in the generated video or in the evaluator's judgment of it. Over 450 videos and 375 comparisons per benchmark (900 videos in all), rewriting changes the physical score in 35.5% and 35.7% of comparisons, flips the official VideoPhy-2 pass decision in 13.3%, and changes the point ranking of 4 of the 5 models on each benchmark, with 70% of pairwise model orderings preserved. The differences that a single-prompt ranking reports are within the variation caused by the choice of prompt, so the conclusions of existing benchmarks depend on the prompt, not on the models alone.

To address this problem, we propose *Prompt Set Evaluation*. Instead of a single prompt, each scenario is represented by a set of K validated equivalent prompts, shared across all models, and a model is scored by its mean performance over the set together with the variation across prompts, so that the evaluation covers the distribution of valid descriptions of a scenario rather than one of them. Because the protocol multiplies the evaluation cost by K, we also study how reliability depends on K, where reliability means reproducibility across equivalent prompts, not validity against true physical ability. Two single-prompt evaluations that use different equivalent prompts agree on the ordering of two models in only 49% and 61% of cases; with five prompts per scenario, scores and rankings become considerably more stable, and the reported variation exposes the sensitivity of each scenario.

The source of this sensitivity matters as much as its existence. A rewrite may lead the video generation model to realize the intended event less faithfully, in which case the evaluator correctly exposes a weakness of the model; or the videos may realize the same physical event equally well and still receive different results, in which case the evaluator does not measure physical understanding alone. We separate these possibilities by scoring identical videos repeatedly, replacing hosted evaluator components with a fully repeatable local model, and collecting blinded human judgments of physical plausibility. The sensitivity is not random noise, as it persists under fully repeatable evaluators. Shown

pairs of videos whose physical scores differ, human raters rarely share the evaluator's preference and often consider the two videos equally plausible, so the evaluator's score differences often lack a counterpart in human judgments, which suggests that the evaluators also respond to differences unrelated to physics.

Our contributions are as follows:

- We show that physics benchmarks for video generation are sensitive to the prompt: equivalent rewrites change scores, decisions, and the point ranking of five models on two benchmarks, even with fixed generation seeds and physical judgments that never see the rewritten prompt.
- We propose Prompt Set Evaluation, which scores each scenario over a set of validated equivalent prompts and reports mean performance with its variation; we analyze how many prompts are needed and check the scores on an independently constructed prompt set.
- We analyze the source of the sensitivity with repeated scoring, a fully repeatable local evaluator, and a blinded human study, and find that it is not evaluator noise and that human raters rarely share the evaluator's preference behind a score change.

2 Related Work

Physical evaluation of video generation. General video benchmarks measure visual quality, temporal consistency, and text alignment (Huang et al., 2024; 2026; Liu et al., 2024; 2023). The physics benchmarks we study describe a scenario in text, generate a video, and judge its physical plausibility with a vision-language model or a multi-stage pipeline built on one (Bansal et al., 2025; 2026; Meng et al., 2025; Gu et al., 2026); all of them represent each scenario by a single prompt, whose reliability we examine. Other paradigms do not rely on such judgments. One measures generated motion against physical ground truth: continuations of real videos (Motamed et al., 2026), conservation laws in controlled physical systems (Tragoudaras et al., 2026), or trajectories of falling objects (Li et al., 2025a). Another trains video models on simulated dynamics and tests whether they extrapolate the governing law beyond the training distribution (Kang et al., 2025), and concurrent work evaluates the distribution of outcomes over repeated rollouts (Pu et al., 2026). A third probes video models without generation, through violation-of-expectation tests of intuitive physics (Riochet et al., 2022; Garrido et al., 2025; Bordes et al., 2025) or physical prediction (Bear et al., 2021). These paradigms are complementary to the one we study and do not depend on a text prompt.

Prompt sensitivity and automatic evaluators. Language-model evaluations are sensitive to prompt formatting (Sclar et al., 2024; Mizrahi et al., 2024) and perturbed prompts (Zhu et al., 2024), and text-to-image models to small prompt changes (Du et al., 2023); prompt adaptation and ensembling have been used to improve generation and recognition (Hao et al., 2023; Cheng et al., 2025; Xue et al., 2025; Gu et al., 2025; Radford et al., 2021). We instead measure the prompt sensitivity of physics benchmarks and use validated equivalent prompts to make the evaluation itself more reliable. Automatic video evaluators raise questions of validity (He et al., 2024; Wu et al., 2024), and model-based judges exhibit systematic biases (Zheng et al., 2023; Chen et al., 2024; Shi et al., 2025); we test whether existing evaluators respond to factors other than physics.

3 Prompt Set Evaluation

3.1 From a Single Prompt to a Distribution of Descriptions

A physics benchmark consists of scenarios s, each specified by a text prompt p_0. For a video generation model G_m and the benchmark's evaluator E, the benchmark reports the evaluation result $Y_m(s, p_0) = E(G_m(p_0))$, normalized to $[0, 1]$ so that larger values indicate better physical plausibility; we call this quantity the physical score. Let Q_s denote the distribution of descriptions of scenario s induced by the rewriting procedure of this paper: the instruction of Appendix A applied to p_0 by a language model, followed by the validation step below. The benchmark reports the result at the one fixed description p_0, whereas Prompt Set Evaluation targets

$$\theta_m(s) = \mathbb{E}_{p \sim Q_s}[Y_m(s, p)]\,, \tag{1}$$

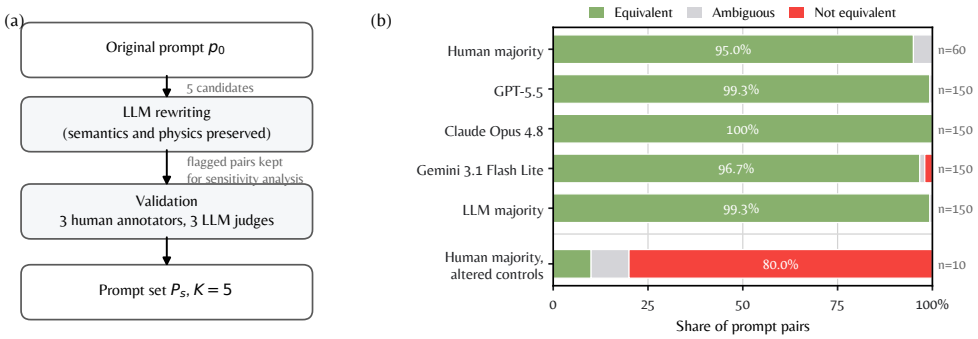


Figure 2: Construction and validation of prompt sets. (a) Each original prompt is rewritten by a large language model under constraints that preserve its semantics and physics, and every rewrite is validated before use. (b) Equivalence judgments. Three annotators judged 60 sampled pairs by majority vote; three language-model judges judged all 150 pairs. The bottom row shows the annotators on 10 control pairs in which a physical detail was altered, which they reject.

the expected physical score under that rewriting procedure, not over all conceivable descriptions of the scenario. Given a prompt set $P_s = \{p_1, \ldots, p_K\}$ sampled from Q_s, the estimate is

$$\widehat{\theta}_m(s) = \frac{1}{K} \sum_{k=1}^{K} Y_m(s, p_k), \tag{2}$$

and the spread of $\{Y_m(s, p_k)\}_k$ measures the prompt sensitivity of the scenario. In our experiments the evaluated pool of a scenario is the original prompt and its five rewrites; the original is a fixed reference rather than a draw from Q_s, so the six-prompt mean weights it as one of six descriptions. Model-level results average $\widehat{\theta}_m(s)$ over scenarios with equal weight, and the same prompt sets are used for every model, so all models are compared under identical descriptions.

3.2 Constructing and Validating Prompt Sets

We generate candidate rewrites of p_0 with a large language model under an instruction that permits only grammatical restructuring, voice changes, and conservative synonym substitution, and forbids adding or removing objects, quantities, actions, directions, materials, forces, camera instructions, or scene details (Appendix A). Human annotators and separately queried language-model judges then decide whether each rewrite describes the same physical scenario as p_0 (Figure 2a). Construction and validation never see a video or an evaluation result, so a prompt set cannot favor any model.

For our prompt sets (Figure 2b), annotators accept 95.0% of sampled pairs and reject none, the language-model judges accept 96.7–100%, and annotators reject 80% of altered controls.

3.3 Measuring Prompt Sensitivity

For a rewrite p_k we define its effect on the physical score as $\delta_k = Y_m(s, p_k) - Y_m(s, p_0)$, with the generation seed held fixed so that the two videos differ only through the prompt. We report the fraction of rewrites with $\delta_k \neq 0$, the mean $|\delta_k|$, and the fraction that change the benchmark's categorical decision. For VideoPhy-2 the physical score is the physical-commonsense (PC) rating on $[1, 5]$ and the decision the official criterion $\mathrm{PC} \geq 4$ and $\mathrm{SA} \geq 4$, where SA is semantic adherence; we also report PC and SA changes separately. For PhyGenBench, the physical score is the final score on $[0, 3]$; as the benchmark defines no decision threshold, we use $\mathrm{final} \geq 2$ as a diagnostic one. For model comparisons we report ranks with paired scenario-level differences and their bootstrap intervals.

3.4 Attributing the Sensitivity

A nonzero δ_k can arise in the video generation model, if the two videos differ in how faithfully they realize the event, or in the evaluator, if it scores physically equivalent videos differently. Scoring an

Table 1: Prompt sensitivity of physics benchmarks. Each model contributes 75 original-to-rewrite comparisons per benchmark. Changed is the fraction of rewrites that change the physical score and Mean $|\delta|$ the mean absolute change as a percentage of the score range. PC, SA, and Joint give the fraction of rewrites that cross the respective VideoPhy-2 threshold; Flip uses the PhyGenBench diagnostic threshold $\mathrm{final} \geq 2$. Bootstrap intervals are in Appendix D.

	VideoPhy-2 (%)					PhyGenBench (%)		
Model	Changed	Mean $\lvert\delta\rvert$	PC	SA	Joint	Changed	Mean $\lvert\delta\rvert$	Flip
CogVideoX-5B	53.3	14.0	22.7	10.7	17.3	37.3	12.4	24.0
SDXL+SVD-XT	40.0	10.3	20.0	12.0	13.3	38.7	12.9	18.7
Cosmos-Predict1-7B	28.0	7.7	10.7	1.3	2.7	42.7	14.7	28.0
Wan2.1-14B	26.7	7.0	17.3	18.7	20.0	24.0	9.8	18.7
HunyuanVideo-13B	29.3	9.0	17.3	5.3	13.3	36.0	12.9	22.7
All models	35.5	9.6	17.6	9.6	13.3	35.7	12.5	22.4

identical video repeatedly measures the evaluator's variability, and replacing hosted evaluator components with a local model under greedy decoding removes it. Decomposing VideoPhy-2 decisions into PC and SA changes, and PhyGenBench scores into their stages, shows which component moves. Blinded human comparisons test whether videos whose physical score changed differ in a way people can see, and generating each prompt with several seeds compares rewriting with sampling.

4 Experiments

4.1 Setup

We evaluate on VideoPhy-2 (Bansal et al., 2026) and PhyGenBench (Meng et al., 2025) with 15 scenarios from each, stratified over their physical categories (listed in Appendix B), and the validated prompt sets of Section 3 (five rewrites per scenario). Five video generation models are tested with their official inference settings (Appendix B): CogVideoX-5B (Yang et al., 2025), SDXL+SVD-XT (Podell et al., 2024; Blattmann et al., 2023a), Cosmos-Predict1-7B (NVIDIA et al., 2025), Wan2.1-14B (Wan Team et al., 2025), and HunyuanVideo-13B (Kong et al., 2024). Unless stated otherwise, each model uses a fixed generation seed shared by all prompts of a scenario. Both evaluators are used exactly as released, except that for two PhyGenBench scenarios Stage 2 uses the compact prompt variant (Appendix B), which does not change the conclusions (Table S13). VideoPhy-2 rates physical commonsense (PC) with a fixed question that does not contain the prompt and semantic adherence (SA) against the candidate caption, and passes a video if both ratings reach 4. PhyGenBench scores each video in three stages from questions and metadata attached to the original scenario. The rewritten prompt thus never enters the physical judgment of either benchmark and enters only the VideoPhy-2 SA rating. Confidence intervals are 95% bootstrap intervals over scenarios (Appendix C).

4.2 Existing Benchmarks Are Sensitive to the Prompt

We first evaluate every model on the original prompt and its five rewrites, 450 videos and 375 original-to-rewrite comparisons per benchmark (900 videos in total), and measure how the results change.

Scores and decisions. On both benchmarks (Table 1), rewriting changes the physical score in about a third of the comparisons (35.5% and 35.7%), with a mean absolute change of about 10% of the score range, and every model is affected, with changed fractions between 24.0% and 53.3%. The changes go in both directions (Figure 3): rewrites raise the score about as often as they lower it, so the effect largely cancels in aggregate means and is visible only at the scenario level. Categorical conclusions change as well: 13.3% of rewrites flip the official VideoPhy-2 pass decision, and 22.4% cross the PhyGenBench threshold. The changes are not confined to videos that fail to depict the scenario: among the 138 VideoPhy-2 comparisons in which both videos adhere to the caption ($\mathrm{SA} \geq 4$), the PC score still changes in 34.8% and flips the official decision on its own in 13.8%. Nor are they evaluator noise: the VideoPhy-2 evaluator is fully repeatable, and on score-changed pairs the repeat variation of the hosted PhyGenBench stages is about one eighth of the rewrite effect (Section 4.4). Rescoring the videos in one fixed environment reproduces these results (Appendix D.3).

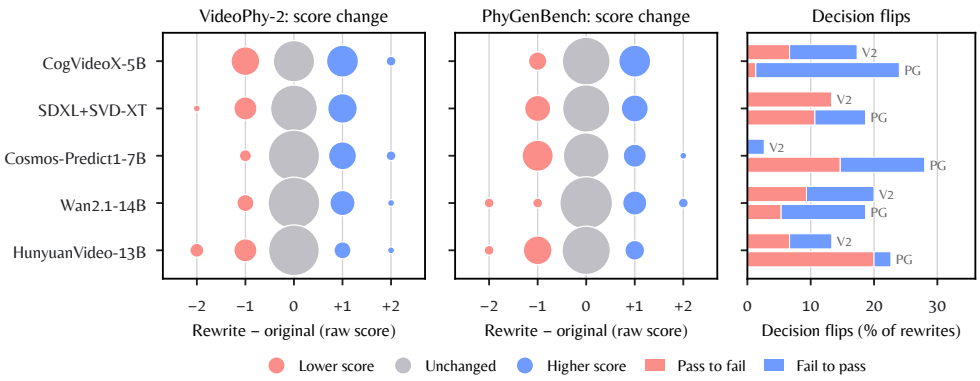


Figure 3: Effect of rewriting on evaluation results. Left and center: change of the raw physical score between each rewrite and the original prompt; circle area is proportional to the fraction of the 75 rewrites per model. Right: fraction of rewrites that flip the decision in each direction (V2: VideoPhy-2 pass decision; PG: PhyGenBench diagnostic threshold).

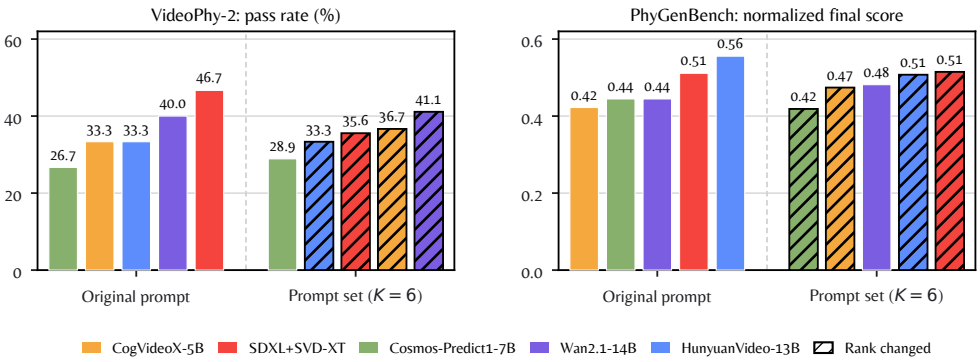


Figure 4: Model rankings under the original prompt alone and under Prompt Set Evaluation with all six prompts. Models are ordered by score within each group; VideoPhy-2 uses the pass rate and PhyGenBench the normalized final score. Hatching marks models whose rank changes.

Rankings. The sensitivity propagates to rankings. Figure 4 ranks the five models by the original prompt alone and by the full prompt set: 4 of the 5 ranks change on each benchmark, and 14 of the 20 pairwise orderings, or 70%, are preserved. The rank changes are not statistically resolved: of the 10 pairwise differences per benchmark, none is resolved under the original prompt, and under the full prompt set 1 is resolved on VideoPhy-2 and 2 on PhyGenBench (Table S8). A single-prompt benchmark thus publishes point ranks without this uncertainty, and the rank changes caused by rewriting occur among models that no single-prompt evaluation separates (Madaan et al., 2024).

4.3 Prompt Set Evaluation and the Set Size

Prompt Set Evaluation replaces the result at one prompt with the mean over a set of K equivalent prompts and reports its spread, at K times the generation and evaluation cost. To determine K, we treat the six observed prompts of each scenario (the original and its five rewrites) as the empirical pool of descriptions of the scenario and resample K of them with replacement. This yields, as a function of K, the standard error of the scenario score and the agreement between the model rankings of two independent K-prompt evaluations (Figure 5); neither uses a reference set, so no K is favored.

With a single prompt, the standard error of a scenario score is 8.8% of the score range on VideoPhy-2 and 12.1% on PhyGenBench, and two independent single-prompt evaluations order a model pair identically in only 49% and 61% of cases. At $K = 5$ the standard error falls to 3.9% and 5.4%, as the $1/\sqrt{K}$ decrease implied by Equation 2, and the agreement reaches 78% and 83%; the remaining

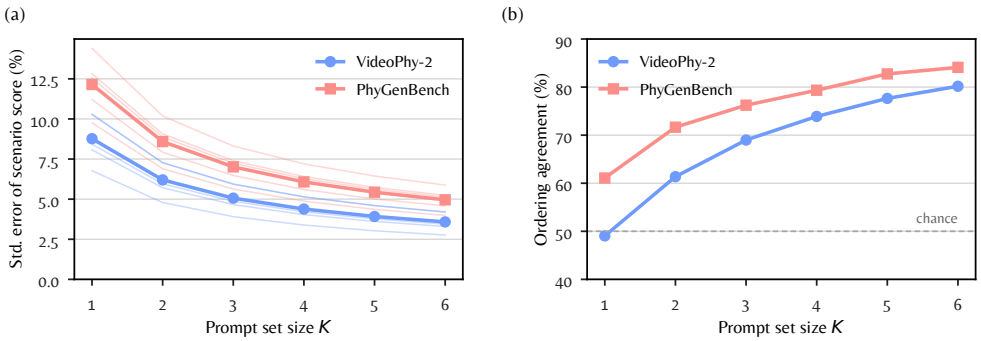


Figure 5: Effect of the prompt set size K. (a) Standard error of the scenario-level physical score, in percent of the score range, averaged over scenarios and models; thin lines are individual models. (b) Agreement of pairwise model orderings between two independent evaluations that each use K prompts per scenario, averaged over 10,000 draws. Agreement uses the sign of the score difference, so a tie agrees only with a tie. The dashed line is the agreement of two random orderings without ties, a reference rather than a chance level under ties.

disagreement concerns model pairs separated by small margins (Appendix E). A sixth prompt lowers the standard error only to 3.6% and 5.0% and raises the agreement only to 80% and 84% for a fifth more generation and evaluation cost, so we use five prompts as the default budget; evaluations on disjoint prompt triples also agree better than single-prompt ones (Appendix E).

An independently constructed prompt set reproduces the aggregate scores. Resampling reuses the prompts of one construction. We therefore built a second set of five rewrites per scenario with the same rewriting instruction and language-model judges, generated without access to the first set and selected to avoid near-copies of it, and evaluated CogVideoX-5B and Wan2.1-14B on it with their main-study seeds (Appendix F). The new rewrites again change the physical score in 25% to 43% of comparisons. Averaged over five prompts, the two sets give benchmark scores within 3% of the score range; scenario scores differ by 3.7% to 10.2%, 20% to 62% less than single prompts, and in five of six model–metric combinations by no more than a random split of the ten prompts, consistent with the second set being another sample from the same distribution. Neither set resolves the order of the two models. The uncertainty estimated from the first set also predicts the second: at the benchmark level, all six model–metric differences lie within one predicted standard deviation, and at the scenario level the 95% prediction interval covers 88% of scenarios, close to the 93% obtained when the ten prompts of each scenario are split into two halves at random (Appendix F).

4.4 Where Does the Sensitivity Come From?

Physical and semantic judgments move independently. If rewriting affected the evaluation only through semantic adherence, the physical judgment would be unchanged. Table 1 shows the opposite: on VideoPhy-2, PC threshold crossings (17.6%) are more frequent than SA crossings (9.6%), and of the 50 joint-decision flips, 19 involve PC alone, 27 SA alone, and 4 both (Appendix D). The flips come from the videos, not from the caption in the SA query. Because that query contains the caption, we rescored the SA of every rewrite video with the original caption (with the rescored scores of Appendix D.3): the joint decision then flips in 15.7% instead of 15.2% of comparisons, and 25 of the 26 SA-only flips remain (Appendix D.4). The PC score should still not be read alone: nearly uniform CogVideoX-5B videos that show no event receive the maximum PC score and are rejected only by the joint decision; excluding them leaves the results above essentially unchanged (Appendix H.2).

Evaluator noise does not explain the sensitivity. We scored all 450 VideoPhy-2 videos and 135 PhyGenBench videos five times with identical inputs (Appendix G.1; Table S11). The VideoPhy-2 evaluator is fully repeatable: across 2,250 evaluations, no PC or SA score changed, so its sensitivity to rewriting is entirely a response to differences between the videos. On PhyGenBench, the local VQAScore stage (Lin et al., 2024) is likewise repeatable, whereas the two hosted GPT-4o stages are not: the final score changed across the five repeats for 18.9% of videos and the decision for 11.8%,

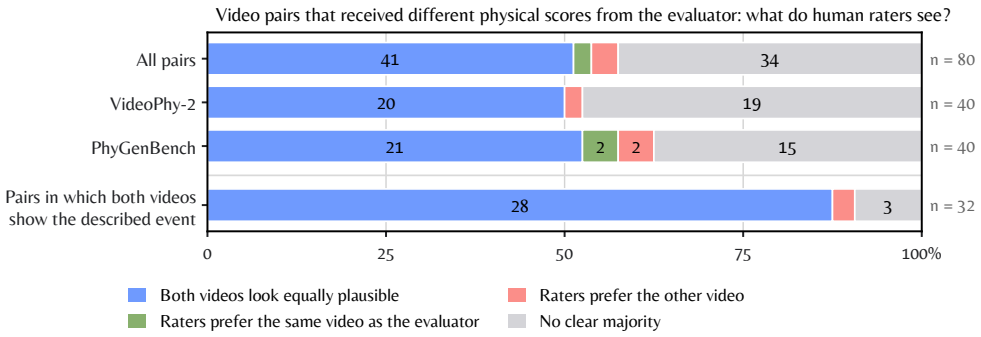


Figure 6: Human judgments of 80 randomly sampled video pairs to which the evaluator assigned different physical scores. Bars count the pairs that at least three of the four raters judged in each way. The bottom bar keeps pairs in which a majority considers both videos to show the event.

mostly through the video-level stage, which is also the stage that varies most across the six videos of a scenario for every model (Appendix G.2). Even there, on the 40 score-changed pairs per benchmark of the repeat study, the change caused by rewriting is about eight times the repeat variation of the same videos (35.0% against 4.2% of the score range). Repeatability does not imply that an evaluator responds only to physics, but the sensitivity of Section 4.2 is not evaluator noise.

The sensitivity persists under a fully repeatable evaluator. To test whether the variation of the hosted stages inflates the PhyGenBench results, we replace GPT-4o in Stages 2 and 3 with a local vision-language model, Qwen3.6-27B with greedy decoding, keeping frames, questions, and scoring rules unchanged (Appendix G.3). This evaluator is fully repeatable: five evaluations of 134 videos never differ in any stage, whereas GPT-4o changes the final score of 19% of the videos scored completely by both. Yet rewriting still changes its final score in 31.4% of the 360 eligible comparisons and its decision in 18.9%, against 36.1% and 23.1% for GPT-4o, and the two evaluators agree on the direction of the changes they both register (Appendix G.3). Prompt sensitivity thus remains when the evaluator is deterministic and when it is exchanged.

Human raters rarely share the evaluator's preference. If rewriting led the models to generate physically different videos, people should prefer the higher-scored video. Four raters not involved in this work compared 80 random pairs of videos from an original prompt and one of its rewrites whose physical scores differ (40 per benchmark). Without seeing the model, the prompts, or the scores, they judged whether each video shows the described event and which is physically more plausible (Figure 6). A majority of at least three raters prefers the video with the higher score in only 2 of the 80 pairs and the other video in 3, judges the two videos equally plausible in 41, and reaches no majority in 34; among the 32 pairs in which both videos show the event, 28 are judged equally plausible. This is not limited to small or unstable score differences: none of the 8 pairs whose scores differ by two or more points and 1 of the 46 pairs whose difference persists when the evaluator rescores the videos receive a majority for the higher-scored video (Appendix G.4). The raters recognized the incorrect video in all six control pairs, animations in which one video gains energy or violates momentum conservation. Leaving out any one rater, equally plausible remains the most frequent judgment.

Summary. The sensitivity is not evaluator noise, as it appears with fully repeatable evaluators. Raters seldom share the evaluator's preference and judge about half of the sampled pairs equally plausible, which suggests that the evaluators also respond to differences unrelated to physics. Prompt sets therefore stabilize the evaluation without making the evaluator more faithful to physics.

4.5 Prompts versus Generation Seeds

To compare the prompt with the generation seed, we run a crossed experiment on VideoPhy-2 with CogVideoX-5B and Wan2.1-14B: for five random scenarios, each of the six prompts is generated with the same five new seeds (150 videos per model; Appendix H). Changing the prompt at a fixed seed changes the PC score by 7.5% and 3.2% of the score range, and changing the seed at a fixed prompt

by 8.3% and 4.5%; the paired intervals of their difference include zero, and their order reverses for CogVideoX-5B without degenerate generations (Appendix H.2). Rewording a prompt thus changes the physical score as much as resampling the video, although only the latter is usually treated as a source of variation (Bouthillier et al., 2021; Agarwal et al., 2021). At a fixed seed, a rewrite changes the video realized from the same noise; permuting the prompts within each seed shows that for CogVideoX-5B the prompt also changes the expected PC score beyond seed variation ($p = 0.004$), whereas for Wan2.1-14B, whose scores vary less, the seed-averaged scores of different prompts stay within seed variation, and for neither model does the original prompt differ systematically from its rewrites (Appendix H.1). With five generations per scenario, five prompts with different seeds estimate the PC score with a lower error than one prompt with five seeds for CogVideoX-5B (3.8% against 5.3% of the range) and with a similar error for Wan2.1-14B (2.0% and 2.0%; Table S15), so spreading the generations over prompts and seeds covers both sources at the same cost. On the joint decision, which is constant within most of these scenarios, no allocation is resolved (Appendix H).

5 Discussion and Limitations

Evaluate scenarios, not prompts. A single-prompt score reflects the model's behavior under that description and does not transfer to equivalent ones. Prompt Set Evaluation instead measures expected performance over valid descriptions and reports its variation, without changing the benchmark's scenarios, criteria, or evaluator, and makes results reproducible across equivalent prompts. Five prompts per scenario substantially improve stability, and a different seed per prompt also covers sampling variation at no extra cost. At a fixed total budget, prompts per scenario trade against the number of scenarios; five per scenario is the choice when the scenarios are fixed. Benchmarks should release validated prompt sets and report the spread across prompts with the mean.

Evaluator scores and human judgments diverge. The sensitivity is not evaluator noise, although hosted evaluator stages flip about one decision in ten on repeat and should be scored repeatedly. Raters seldom share the evaluator's preference and find about half of the score-changed pairs equally plausible. Rewriting changes a video's framing, objects, and how completely the event is shown; whether evaluators respond to such non-physical differences or to physical differences raters do not perceive can be decided by controlled changes of appearance at fixed physics. Prompt sets average over this variation without removing its cause; evaluators trained to be invariant to non-physical differences are a natural next step.

Limitations. We study one paradigm, text-conditioned generation scored by model-based evaluators; approaches without a text description, such as measuring motion against ground truth or probing representations (Motamed et al., 2026; Tragoudaras et al., 2026; Li et al., 2025a; Kang et al., 2025; Garrido et al., 2025; Bordes et al., 2025), are outside its scope. Its intervals describe a panel of two benchmarks with 15 scenarios each and five models, not benchmark-wide prevalence. Equivalence is judged from the text, and the independent set covers two of the five models. The human study, with four raters, shows that evaluator and human judgments diverge but does not identify the cause. PhyGenBench scores come from one evaluation per video, whose variation the repeat study bounds at one eighth of the rewrite effect on score-changed pairs. Prompt Set Evaluation multiplies the cost by the set size and does not correct an evaluator that misjudges physics.

6 Conclusion

Physics benchmarks for video generation are sensitive to the prompt that describes each scenario: across five models and two benchmarks, equivalent rewrites change the physical score in about a third of the comparisons, flip official decisions, and change the models' point ranking, even with fixed seeds and unchanged evaluators. Prompt Set Evaluation scores each scenario over a validated set of equivalent prompts and reports the mean with its variation; five prompts per scenario substantially improve the stability of scores and rankings, and an independently constructed prompt set reproduces the aggregate scores. The sensitivity is not evaluator noise, and human raters seldom share the evaluator's preference, which suggests that evaluators also respond to differences unrelated to physics. Physical understanding should be evaluated over the distribution of descriptions of a scenario rather than a single prompt.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure. *Data generation:* GPT-5.5 generated the rewritten prompts that constitute the studied intervention, and GPT-5.5, Claude Opus 4.8, and Gemini 3.1 Flash Lite judged their equivalence; GPT-4o and Qwen3.6-27B are components of the evaluators under study. *Methodology, experiments, and implementation:* the authors defined the research questions and the experimental protocols and made all decisions; AI coding assistants helped to elaborate protocol details and statistical analyses, implemented and ran the generation, scoring, and analysis code, cleaned and reconciled experimental records, and assisted in discussing the results. *Translation:* the Chinese translations of the scenario descriptions offered to raters as an aid were produced with AI tools; all raters read English and relied primarily on the original English descriptions. Generative AI tools were not used for theoretical models, mathematical claims, or proofs, which this work does not contain. Additionally, we used generative AI tools for drafting and editing the text of this paper, producing figures, checking the bibliography, and literature search. We have reviewed all AI-assisted work: the authors personally inspected all data, experimental results, and generated videos, reviewed the main experiment and analysis code (with the exception of simple visualization scripts), and checked the text and its claims against the results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates publicly released models and benchmarks. It includes two human studies, in which annotators judged pairs of text prompts and pairs of generated videos. All participants were informed about the purpose and content of the studies and consented to take part; the judgments are stored and will be released in de-identified form and contain no personal information. Prompt rewriting preserves the meaning of benchmark prompts for measurement purposes and is not designed to degrade deployed systems.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A and B specify prompt construction and validation, the models with their inference settings, and the evaluators; Appendix C specifies all statistical procedures, and Appendix G.4 the human study. Upon acceptance, we will release all materials needed to reproduce the results: the prompt sets, equivalence judgments, per-video evaluation results, de-identified human ratings, inference configurations, and the experiment and analysis code. Generated videos will be released where the model licenses permit; model checkpoints are public, and their versions and settings are listed in Appendix B.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

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

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024. URL `https://openai.com/index/video-generation-models-as-world-simulators/`.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning (ICML)*, 2024.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, et al. VPO: Aligning text-to-video generation models with prompt optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Chengbin Du, Yanxi Li, Zhongwei Qiu, and Chang Xu. Stable diffusion is unstable. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Jing Gu, Ashwin Nagarajan, Tejas Polu, Kaizhi Zheng, Ruijian Zha, Jie Yang, and Xin Eric Wang. CCC: Enhancing video generation via structured MLLM feedback. In *ICML Workshop on Test-Time Adaptation: Putting Updates to the Test (PUT)*, 2025.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations (ICLR)*, 2026.

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.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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 (EMNLP)*, 2024.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

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 (CVPR)*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 48(3):3268–3285, 2026.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning (ICML)*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning (ICML)*, pp. 35685–35709, 2025a.

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. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025b.

Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, et al. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, 2024.

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 (CVPR)*, 2024.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning (ICML)*, pp. 43781–43806, 2025.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics (TACL)*, 12:933–949, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Yuandong Pu, Le Zhuo, Sayak Paul, Gabriel Jorge Menezes, Avram Djordjević, Shiyang Li, Yifan Zhou, Bin Fu, Wenlong Zhang, Junjun He, Yu Qiao, Yihao Liu, Jinbo Xing, and Xi Chen. PAWBench: How far are we from probabilistically aligned world modeling? *arXiv preprint arXiv:2608.27345*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(9):5016–5025, 2022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *International Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023.

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating newtonian mechanics in video generative models with real physical systems. In *International Conference on Machine Learning (ICML)*, 2026. URL `https://arxiv.org/abs/2504.02918`.

Wan Team, 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.

Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Sherry Yang, Jacob C. Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making. In *International Conference on Machine Learning (ICML)*, 2024b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts. In *ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis (LAMPS)*, 2024.

SUPPLEMENTARY MATERIAL

A PROMPT CONSTRUCTION AND VALIDATION

For each of the 30 original prompts, GPT-5.5 generated 14 candidate rewrites in one request with the following instruction. Candidates that are empty or fragmentary, longer than 1.6 times the original, add framing words absent from the original, or duplicate the original or an earlier candidate after lowercasing are discarded automatically, and the first five remaining candidates form the prompt set. Generated videos and evaluation results play no role in constructing or filtering the rewrites.

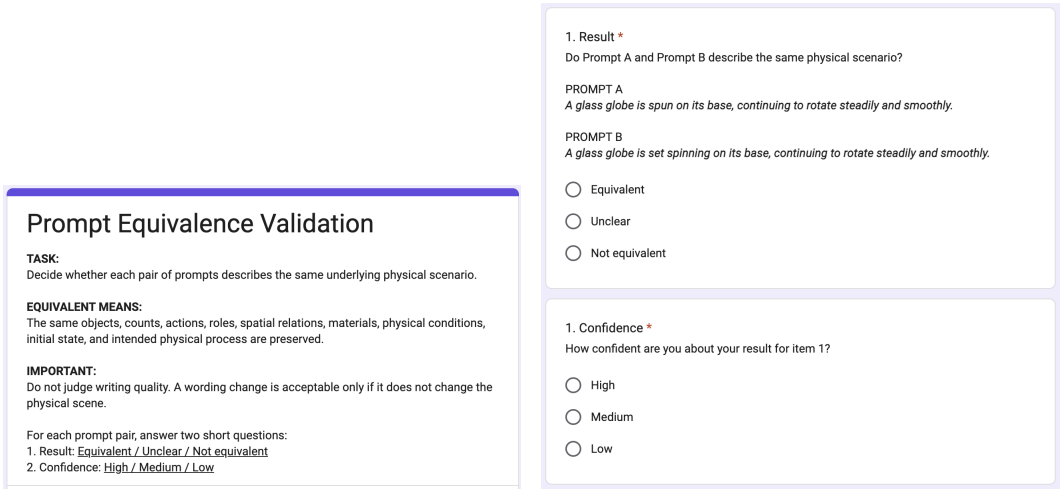
```
Task: Generate {num_return} candidate text prompts that preserve
the exact same physical scene and semantic meaning as the original
prompt.
Allowed changes:
1. Minimal grammatical restructuring, such as passive voice or
phrase reordering.
2. Conservative synonym substitution only when it does not change
entities, materials, quantities, direction, speed, force, or physical
dynamics.
Strict constraints:
- Preserve the same subjects, objects, materials, actions, spatial
relations, and physical phenomena.
- Do not add or remove any object, location, background, camera
detail, visual style, lighting, timing, force, speed, or physical
modifier.
- Do not make the event more detailed, cinematic, or scene-like than
the original.
- Do not add meta-framing phrases unless the original already
contains that framing concept.
- Every candidate must be a complete replacement prompt and remain
close in length to the original.
Original Prompt: {caption}
Return only valid JSON with a top-level "candidates" array of
{num_return} strings.
```

Human validation. Three annotators judged 60 pairs (two rewrites per scenario) and ten negative controls through the interface in Figure S1, deciding whether the rewrite describes the same physical scenario as the original. Majority labels were equivalent for 57 pairs and ambiguous for three (Table S1); pairwise agreement was 88.9%, 51 pairs were unanimous, and eight controls were rejected.

Table S1: Prompt-equivalence judgments. Human labels are majority votes over three annotators on 60 sampled pairs; language-model labels are majority votes over three judges on all 150 pairs.

Judge	Equivalent	Ambiguous	Not equivalent	Total
Human, VideoPhy-2 pairs	29 (96.7%)	1 (3.3%)	0	30
Human, PhyGenBench pairs	28 (93.3%)	2 (6.7%)	0	30
GPT-5.5	149 (99.3%)	1 (0.7%)	0	150
Claude Opus 4.8	150 (100.0%)	0	0	150
Gemini 3.1 Flash Lite	145 (96.7%)	2 (1.3%)	3 (2.0%)	150
Three-judge majority	149 (99.3%)	1 (0.7%)	0	150

Language-model validation. Three judges received the rubric below for all 150 pairs. The single pair not accepted by a majority replaced "pushes" with "shoves." Together with the three ambiguous human labels, this yields four distinct flagged pairs, which the sensitivity analysis in Table S5 excludes.



(a) Task instructions and prompt-pair layout.

(b) Equivalence and confidence response fields.

Figure S1: Interface used for the human prompt-equivalence validation.

```
System: You are validating text prompts for a video-generation
physics benchmark. Judge only whether the rewritten prompt preserves
the same physical-semantic scenario as the original prompt. Do
not judge writing quality, style, or whether the prompt is better.
Return only JSON.

Task: Decide whether the rewritten prompt describes the same
physical-semantic scenario as the original prompt.

Equivalent means all physically important details are preserved:
objects, counts, actions, roles, spatial relations, materials,
physical conditions, initial/final state, direction, force, speed,
and intended physical process.
Ambiguous means the pair is mostly similar but a phrase could
plausibly imply a physical change.
Not equivalent means the rewrite changes at least one physically
important detail.

Pair ID: {pair_id}
Benchmark: {benchmark}
Original prompt: {original_prompt}
Rewritten prompt: {rewritten_prompt}

Return only JSON with fields: decision, confidence,
changed_dimensions, and short_reason.
```

B Generation and Evaluation Protocol

Table S2: Inference settings and compute for the five video generation models. For SDXL+SVD-XT, the two step counts refer to SDXL image generation and SVD video generation; each HunyuanVideo-13B sample uses four H100s.

Model	Resolution	Frames / FPS	Sampling	Guidance and motion settings	Precision	GPUs
CogVideoX-5B	720×480	49 / 8	50 DDIM steps	CFG $= 6.0$	BF16	$1\times$ A100
SDXL+SVD-XT	1024×576	25 / 7	40 + 25 Euler steps	SDXL CFG $= 7.0$; SVD CFG $1.0 \rightarrow 3.0$; motion bucket 127; noise aug. 0.02	FP16	$1\times$ A100
Cosmos-Predict1-7B	1280×704	121 / 24	35 official steps	CFG $= 7.0$	BF16	$1\times$ A100-40GB / H100-80GB
Wan2.1-14B	832×480	81 / 16	50 UniPC steps	CFG $= 5.0$; flow shift $= 5.0$	BF16	$1\times$ H100-80GB
HunyuanVideo-13B	1280×720	129 / 24	50 flow-matching steps	reverse flow; shift $= 7.0$; embedded CFG $= 6.0$	BF16, FP8 DiT	$4\times$ H100-80GB

Wan2.1-14B and Cosmos-Predict1-7B use their official generation code with prompt expansion disabled; HunyuanVideo-13B uses the official container with sequence parallelism; CogVideoX-5B and SDXL+SVD-XT use Diffusers. SDXL+SVD-XT generates a conditioning image from the

prompt and then animates it. Cosmos-Predict1-7B was run on both A100 and H100 accelerators with identical checkpoints and settings; Table S5 reports the homogeneous-H100 subset as a check.

Scenarios. The 15 scenarios of each benchmark cover every category, one per fine-grained action (VideoPhy-2) or sub-category (PhyGenBench), and were fixed before the prompt sets were built and before any video was generated. VideoPhy-2 (8 Sports and Physical Activities, 7 Object Interactions): `test_00003`, `test_00020`, `test_00022`, `test_00046`, `test_00055`, `test_00057`, `test_00075`, `test_00110`, `test_00408`, `test_00567`, `test_00609`, `test_00702`, `test_00760`, `test_00923`, `test_01648`. PhyGenBench (3 each of Force, Light, Heat, Physical Properties, and Chemical Properties): `phygen_005`, `phygen_016`, `phygen_025`, `phygen_050`, `phygen_075`, `phygen_085`, `phygen_090`, `phygen_096`, `phygen_105`, `phygen_120`, `phygen_125`, `phygen_135`, `phygen_145`, `phygen_150`, `phygen_155`.

Generation seeds. Every model uses one fixed generation seed shared by the six prompts of a scenario, so each comparison contrasts one realization per prompt. The seed is 42 for CogVideoX-5B and SDXL+SVD-XT and 1234 for Cosmos-Predict1-7B, Wan2.1-14B, and HunyuanVideo-13B; the random generator is reset to this seed before every video.

Evaluators. The VideoPhy-2 evaluator is run in evaluation mode with greedy decoding. Its PC question asks whether the video adheres to physical laws and does not contain the caption; its SA question inserts the candidate caption. PhyGenBench resolves each scenario's questions and metadata from the original caption, so the rewritten prompt is never shown to the evaluator. Stage 1 retrieves key frames with CLIP and scores them with VQAScore (`clip-flant5-xxl`) before a fixed bucketing to 0–3; Stages 2 and 3 are GPT-4o judgments of the scenario-specific physical order (first, middle, and last frame) and of the scenario's physics rubric (26 sampled frames). For scenarios `phygen_025` and `phygen_150`, Stage 2 used the compact checks prompt instead of the released prompt variant; Stages 1 and 3 are unchanged, and Table S13 shows that excluding these two scenarios does not change the conclusions. Hosted model identifiers and request settings are fixed across all calls.

C Statistical Procedures

Confidence intervals resample the 15 scenarios of a benchmark with replacement (10,000 draws), keeping all prompts and models of a scenario together, and report the 2.5th and 97.5th percentiles. They describe sensitivity to the composition of the panel rather than a probability sample of the benchmark. Paired model differences are computed within each bootstrap draw. The set-size analysis of Section 4.3 resamples K of the six observed prompts of each scenario with replacement. The standard error in Figure 5a is $\sigma_s/\sqrt{K}$, with σ_s the population standard deviation of the six normalized physical scores of a scenario, averaged over scenarios and models. The agreement in Figure 5b draws, for each of 10,000 repetitions, two independent evaluations in which every scenario receives K resampled prompt indices shared across models, computes each model's official aggregate (VideoPhy-2 pass rate or PhyGenBench normalized final score), and reports the fraction of the ten model pairs ordered identically by the two evaluations. Orderings are compared through the sign of the score difference, so two evaluations agree on a pair when both give the same sign; a tie in one evaluation and not the other counts as a disagreement, and two ties count as an agreement. The 50% line in Figure 5b is the agreement of two independent random orderings without ties. It is a reference value rather than an exact chance level, because the aggregates take few distinct values and ties therefore occur. The split-half check enumerates the ten partitions of the six prompts of a scenario into two disjoint triples for the scenario-level difference, and draws random per-scenario partitions (10,000 repetitions, shared across models) for the ordering agreement; the single-prompt baseline uses two distinct prompts of the same scenario. The finite-pool diagnostic in Appendix E instead draws subsets without replacement, always including the original prompt, and compares them with the full six-prompt set. In the seed experiment, the prompt effect is the mean absolute score difference over all pairs of prompts at the same seed, and the seed effect the mean absolute difference over all pairs of seeds at the same prompt, averaged within a scenario and then over scenarios (Appendix H).

D Additional Results on Prompt Sensitivity

Bootstrap intervals in this section resample scenarios within the fixed 15-scenario panel of each benchmark. They describe sensitivity to the panel's composition and do not include evaluator repetitions.

D.1 Per-Model Results with Bootstrap Intervals

Table S3 gives the per-model values of Table 1 with fixed-panel 95% bootstrap intervals, and Table S4 partitions the VideoPhy-2 joint-decision flips into those caused by PC alone, SA alone, or both.

Table S3: Prompt sensitivity by model with 95% bootstrap intervals. Mean $|\delta|$ and Set range are percentages of the score range; Set range is the range of the six scores of a scenario, averaged over scenarios. Changed and Outcome transition use 75 rewrites per model; Outcome transition is the VideoPhy-2 joint decision or the PhyGenBench diagnostic threshold.

Benchmark	Model	Mean $\|\delta\|$	Set range	Changed	Outcome transition
VideoPhy-2	CogVideoX-5B	14.0 [8.7, 19.3]	25.0 [16.7, 33.3]	53.3 [34.7, 72.0]	17.3 [2.7, 34.7]
	SDXL+SVD-XT	10.3 [6.7, 14.0]	25.0 [16.7, 33.3]	40.0 [26.7, 53.3]	13.3 [0.0, 29.3]
	Cosmos-Predict1-7B	7.7 [3.3, 13.0]	16.7 [10.0, 25.0]	28.0 [13.3, 45.3]	2.7 [0.0, 6.7]
	Wan2.1-14B	7.0 [4.0, 10.3]	23.3 [13.3, 33.3]	26.7 [14.7, 38.7]	20.0 [5.3, 37.3]
	HunyuanVideo-13B	9.0 [4.3, 14.0]	20.0 [10.0, 30.0]	29.3 [13.3, 46.7]	13.3 [2.7, 25.3]
PhyGenBench	CogVideoX-5B	12.4 [7.6, 17.3]	24.4 [17.8, 31.1]	37.3 [22.7, 52.0]	24.0 [9.3, 40.0]
	SDXL+SVD-XT	12.9 [8.0, 18.2]	31.1 [24.4, 37.8]	38.7 [24.0, 54.7]	18.7 [8.0, 32.0]
	Cosmos-Predict1-7B	14.7 [9.3, 20.0]	33.3 [22.2, 44.4]	42.7 [26.7, 58.7]	28.0 [14.7, 42.7]
	Wan2.1-14B	9.8 [4.0, 16.4]	22.2 [11.1, 35.6]	24.0 [10.7, 38.7]	18.7 [5.3, 34.7]
	HunyuanVideo-13B	12.9 [7.6, 18.7]	31.1 [17.8, 46.7]	36.0 [21.3, 52.0]	22.7 [10.7, 34.7]

Table S4: Partition of VideoPhy-2 joint-decision flips. Counts use 75 rewrites per model and 375 in total.

Model	PC flip	SA flip	Joint flip	PC only	SA only	Both
CogVideoX-5B	17	8	13	6	4	3
SDXL+SVD-XT	15	9	10	1	9	0
Cosmos-Predict1-7B	8	1	2	1	1	0
Wan2.1-14B	13	14	15	4	10	1
HunyuanVideo-13B	13	4	10	7	3	0
All models	66	36	50	19	27	4

D.2 Robustness Checks

Table S5 repeats the main measurements after excluding the four flagged prompt pairs and on the two models run exclusively on H100 accelerators.

Table S5: Robustness of the main results. Changed is a nonzero change of the physical score; Flip is a VideoPhy-2 joint-decision flip or a PhyGenBench diagnostic threshold crossing.

Benchmark	Subset	Pairs	Changed	Flip
VideoPhy-2	All models	375	133 (35.5%)	50 (13.3%)
	Flagged pairs excluded	365	127 (34.8%)	48 (13.2%)
	H100-only models	150	42 (28.0%)	25 (16.7%)
	H100-only models, flagged pairs excluded	146	39 (26.7%)	24 (16.4%)
PhyGenBench	All models	375	134 (35.7%)	84 (22.4%)
	Flagged pairs excluded	365	131 (35.9%)	84 (23.0%)
	H100-only models	150	45 (30.0%)	31 (20.7%)
	H100-only models, flagged pairs excluded	146	44 (30.1%)	31 (21.2%)

D.3 Reproduction in a Fixed Evaluation Environment

Every qualitative result of Sections 4.2 and 4.3 holds when the videos are scored again in one fixed evaluation environment (Table S6). All 450 VideoPhy-2 videos were rescored in the environment of the repeat study (Appendix G.1), which uses the same evaluator code and released checkpoint, the same 32-frame sampling and greedy decoding, and the same versions of the deep-learning and video-decoding libraries as the main study; its five repeats give identical scores. The hosted PhyGenBench stages cannot be fixed, so the deterministic local evaluator of Appendix G.3 serves as the fixed PhyGenBench evaluator; it scores 449 of the 450 videos, and it is compared with the GPT-4o scores of the main study (Table 1) on the same 370 comparisons and, for rankings and set size, on the 14 scenarios it scores completely. With the rescored VideoPhy-2 scores, rewriting changes the PC score in 36.5% of comparisons against 35.5% in the main study and flips the joint decision in 15.2% against 13.3%; in both, 4 of the 5 ranks change, 7 of the 10 pairwise orders are preserved, none of the rank changes is statistically resolved, and the standard error falls from one to five prompts by more than half. Under the local PhyGenBench evaluator, rewriting changes the score in 31.4% of comparisons and the decision in 18.6%, 3 of the 5 ranks change with 7 of the 10 pairwise orders preserved (4 and 7 for GPT-4o on the same scenarios), and the set-size gains remain.

Table S6: Main results as reported in Section 4.2 and Figure 5 (Main study) and after rescoring in one fixed evaluation environment (Rescored). VideoPhy-2 uses the same evaluator for both, on all 375 comparisons. PhyGenBench compares the GPT-4o scores of the main study with the deterministic local evaluator of Appendix G.3 on the 370 comparisons valid for both; its ranking and set-size rows use the 14 scenarios that both score completely. Changed, Mean $|\delta|$, crossings, and flips are defined as in Table 1; SE and Agreement are the set-size statistics of Figure 5 at $K = 1$ and $K = 5$.

	VideoPhy-2		PhyGenBench	
	Main study	Rescored	GPT-4o	Local
Physical score changed (%)	35.5	36.5	35.9	31.4
Mean $\|\delta\|$ (% of range)	9.6	10.2	12.6	11.2
PC threshold crossings (%)	17.6	21.3	–	–
SA threshold crossings (%)	9.6	9.3	–	–
Decision flips (%)	13.3	15.2	22.4	18.6
PC changed, both SA ≥ 4 (%)	34.8	34.1	–	–
Joint flips, both SA ≥ 4 (%)	13.8	18.8	–	–
Ranks changed (of 5)	4	4	4	3
Pairwise orders preserved (of 10)	7	7	7	7
SE at $K = 1$ / $K = 5$ (%)	8.8 / 3.9	9.4 / 4.2	12.8 / 5.7	10.8 / 4.8
Agreement at $K = 1$ / $K = 5$ (%)	49 / 78	47 / 76	62 / 84	63 / 82

D.4 Semantic Adherence with a Fixed Caption

The joint-decision flips come from the videos, not from the caption in the SA query. The VideoPhy-2 PC question does not contain the caption, but the SA question does, so an SA change between the original video and a rewrite video could come from the rewrite caption rather than from the video. To separate the two, we scored all 450 VideoPhy-2 videos of the five models again with the same evaluator revision, frame sampling (32 frames), greedy decoding, and prompt templates: PC of every video, SA of the 75 original videos with the original caption, and SA of the 375 rewrite videos twice, with the rewrite caption as in the main study and with the original caption. The videos are those of the main study (Appendix B). All 1,275 queries returned valid scores without retries, and 25 repeated queries gave identical results. These are the rescored scores of Appendix D.3, and all comparisons below use them; with them, rewriting changes the PC score in 36.5% of comparisons (137 of 375), against 35.5% in the main study.

With the original caption, the joint decision flips for 59 of the 375 rewrites (15.7% [7.7, 24.0]); with the rewrite caption, for 57 (15.2% [7.7, 23.2]); the difference is -0.5 points $[-1.9, 0.5]$. With the rewrite caption, 26 flips involve PC alone, 26 SA alone, and 5 both; with the original caption, 26, 28, and 5. Of the 26 SA-only flips under the rewrite caption, 25 remain flips under the original caption. SA crosses the threshold for 35 rewrites (9.3%) with the rewrite caption and for 37 (9.9%) with the original caption, and differs from the original video for 94 (25.1%) and 86 (22.9%). The caption

Table S7: Semantic adherence of the rewrite videos with the rewrite caption and with the original caption (75 rewrites per model, 375 in total). Joint flips counts joint-decision flips relative to the original video. SA-only flips remaining counts the SA-only flips under the rewrite caption that remain joint flips under the original caption. The last two columns compare the two captions on the same rewrite video: SA differs is the number of videos whose SA score differs, and SA crosses the number whose SA decision differs.

	Joint flips			Between captions	
Model	Rewrite caption	Original caption	SA-only flips remaining	SA differs	SA crosses
CogVideoX-5B	13	13	4 of 4	7	0
SDXL+SVD-XT	10	10	9 of 9	8	0
Cosmos-Predict1-7B	9	9	0 of 1	6	2
Wan2.1-14B	16	17	10 of 10	6	3
HunyuanVideo-13B	9	10	2 of 2	5	1
All models	57	59	25 of 26	32	6

alone changes the SA score of a rewrite video for 32 of the 375 videos (8.5% [4.0, 14.4]) and moves it across the threshold for 6 (1.6% [0.5, 2.9]). Table S7 gives the counts per model. Fixing the caption thus leaves the decision changes in place: they come from the videos, not from the caption in the SA query.

D.5 Pairwise Model Differences under the Two Protocols

Table S8 reports, for all 10 model pairs of each benchmark, the paired difference of scenario-level scores under the original prompt alone and under the full six-prompt set, with 95% scenario-cluster bootstrap intervals (15 scenarios resampled with replacement, 10,000 draws). A scenario-level score is the mean over the prompts used, and the model aggregate is the official one, the VideoPhy-2 pass rate or the PhyGenBench normalized final score. Under the original prompt, no pair is statistically resolved on either benchmark; under the full prompt set, 1 pair is resolved on VideoPhy-2 and 2 on PhyGenBench. The point order differs between the two protocols for 3 of the 10 pairs on each benchmark, so 70% of the pairwise orderings are preserved; 2 of these 6 pairs are exact ties under the original prompt. In every pair whose order differs, both intervals contain zero, so none of the rank changes of Figure 4 is statistically resolved.

D.6 Dispersion and Mean Shift

Let y_{sk} be the normalized physical score of scenario s for prompt k, with $k = 0$ the original and $k = 1, \ldots, 5$ the rewrites. We define

$$\mathrm{NPV}_s = 4\,\mathrm{Var}_k(y_{sk}), \qquad \mathrm{NMS}_s = \frac{1}{5}\sum_{k=1}^{5} y_{sk} - y_{s0}, \qquad \mathrm{ANMS}_s = |\mathrm{NMS}_s|,$$

where the factor four normalizes the maximum variance on $[0, 1]$ to one. NMS is the signed shift of the rewrite mean from the original score, and ANMS is its magnitude. Table S9 averages these over scenarios. Signed shifts are small while dispersion is substantial, which confirms that positive and negative changes cancel in aggregate means.

D.7 Scenario-Level Score Distributions

Figure S2 shows the physical score of every video by scenario, model, and benchmark. Rewrites frequently straddle the original prompt's score on both sides, and the scenarios with the largest spread differ between models, so no single scenario or model drives the aggregate sensitivity.

E Prompt Set Size

The split-half check of Section 4.3 compares two evaluations that use disjoint halves of each scenario's prompts (three prompts each). They differ in scenario score by 6.4% and 8.5% of the score range

Table S8: Paired scenario-level differences between models under the original prompt alone and under the full six-prompt set, in percentage points of the model aggregate (VideoPhy-2 pass rate; PhyGenBench normalized final score), with 95% scenario-cluster bootstrap intervals. A dagger (†) marks a difference that is statistically resolved, that is, whose interval excludes zero. An asterisk (∗) marks a pair whose point order differs between the two protocols. Models are abbreviated CogVideoX-5B (Cog), SDXL+SVD-XT (SVD), Cosmos-Predict1-7B (Cos), Wan2.1-14B (Wan), and HunyuanVideo-13B (Hun).

		Original prompt		Full prompt set	
Benchmark	Pair	Δ	95% CI	Δ	95% CI
VideoPhy-2	Cog − Cos	6.7	[−20.0, 33.3]	7.8	[−7.8, 26.7]
	Cog − Hun*	0.0	[−26.7, 26.7]	3.3	[−11.1, 20.0]
	Cog − SVD*	−13.3	[−40.0, 13.3]	1.1	[−8.9, 10.0]
	Cog − Wan	−6.7	[−33.3, 20.0]	−4.4	[−14.4, 4.4]
	Cos − Hun	−6.7	[−20.0, 0.0]	−4.4	[−16.7, 6.7]
	Cos − SVD	−20.0	[−40.0, 0.0]	−6.7	[−22.2, 5.6]
	Cos − Wan	−13.3	[−33.3, 0.0]	−12.2†	[−25.6, −1.1]
	Hun − SVD	−13.3	[−40.0, 13.3]	−2.2	[−15.6, 11.1]
	Hun − Wan	−6.7	[−20.0, 0.0]	−7.8	[−17.8, 0.0]
	SVD − Wan*	6.7	[−20.0, 33.3]	−5.6	[−15.6, 4.4]
PhyGenBench	Cog − Cos*	−2.2	[−13.3, 8.9]	5.6	[−3.3, 14.4]
	Cog − Hun	−13.3	[−26.7, 0.0]	−3.3	[−13.7, 5.9]
	Cog − SVD	−8.9	[−24.4, 6.7]	−4.1	[−14.1, 5.2]
	Cog − Wan	−2.2	[−15.6, 13.3]	−0.7	[−11.1, 9.6]
	Cos − Hun	−11.1	[−22.2, 0.0]	−8.9†	[−17.4, −1.5]
	Cos − SVD	−6.7	[−17.8, 4.4]	−9.6†	[−15.6, −4.1]
	Cos − Wan*	0.0	[−15.6, 15.6]	−6.3	[−17.0, 4.1]
	Hun − SVD*	4.4	[−6.7, 15.6]	−0.7	[−10.0, 8.9]
	Hun − Wan	11.1	[0.0, 24.4]	2.6	[−5.2, 11.1]
	SVD − Wan	6.7	[−6.7, 20.0]	3.3	[−7.8, 14.4]

Table S9: Dispersion (NPV), signed mean shift (NMS), and absolute mean shift (ANMS) of the physical score across each prompt set, averaged over scenarios.

Benchmark	Model	NPV	NMS	ANMS
VideoPhy-2	CogVideoX-5B	0.057	0.027	0.120
	SDXL+SVD-XT	0.056	0.017	0.083
	Cosmos-Predict1-7B	0.032	0.057	0.077
	Wan2.1-14B	0.045	0.030	0.050
	HunyuanVideo-13B	0.053	-0.037	0.090
PhyGenBench	CogVideoX-5B	0.069	0.062	0.124
	SDXL+SVD-XT	0.080	0.004	0.120
	Cosmos-Predict1-7B	0.113	-0.031	0.120
	Wan2.1-14B	0.081	0.044	0.098
	HunyuanVideo-13B	0.103	-0.058	0.111

on VideoPhy-2 and PhyGenBench and agree on 67% and 74% of pairwise model orderings, against 9.8% and 13.5% and 45% and 57% for two evaluations that each use a different single prompt.

Figure S3 shows how each model's score changes as the prompt set grows, and Figure S4 the complementary finite-pool diagnostic: subsets of size K that always include the original prompt, drawn without replacement, and compared with the full six-prompt set. In that diagnostic, agreement at $K = 6$ is exact by construction, so it measures the cost of omitting observed rewrites rather than the uncertainty of the estimate, which Figure 5 addresses. Several adjacent models are separated by margins smaller than the range of subset means, which is why rankings converge more slowly than scores.

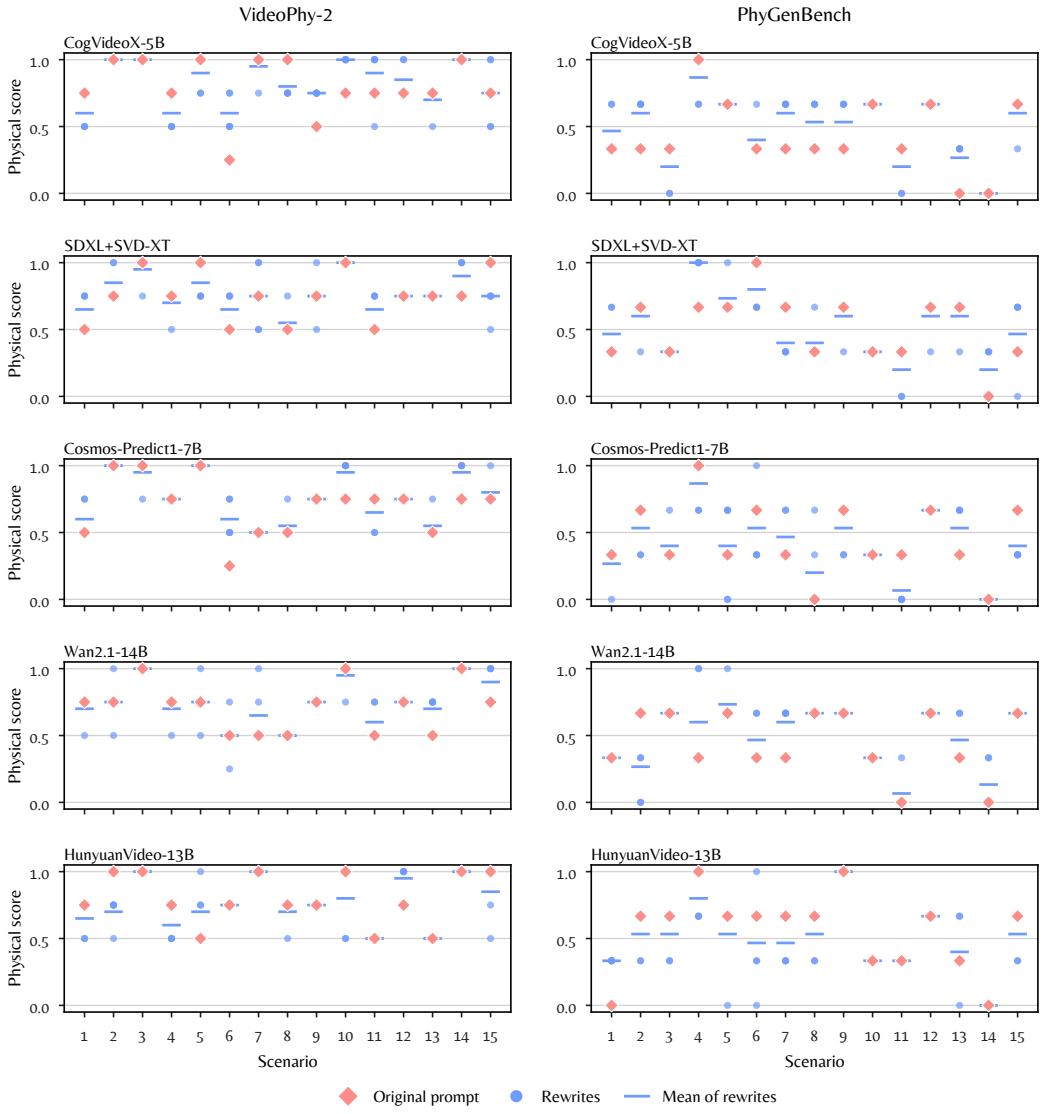


Figure S2: Physical scores of all 900 videos by scenario. Diamonds are the original prompts, circles the five rewrites, and horizontal lines the mean of the rewrites. Scenarios are numbered within each benchmark.

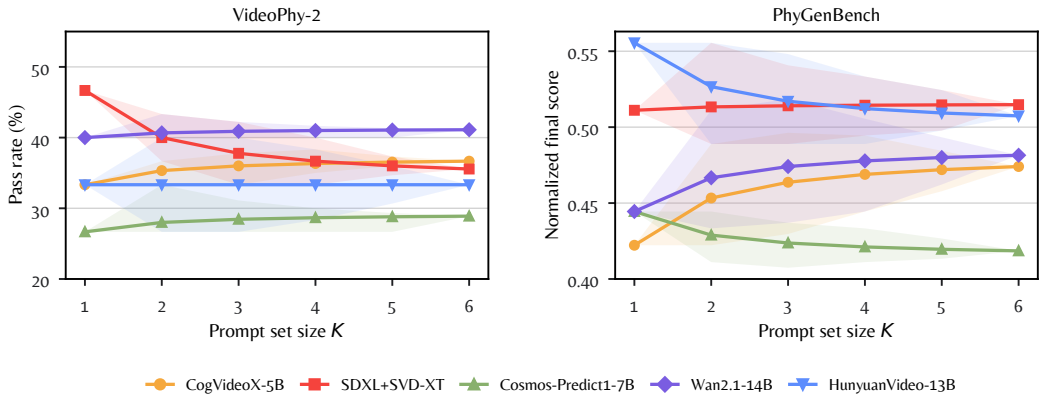


Figure S3: Model scores as a function of the prompt set size K. Lines are the means over all subsets of size K that include the original prompt, and bands span the minimum and maximum over those subsets. Axes are independently scaled.

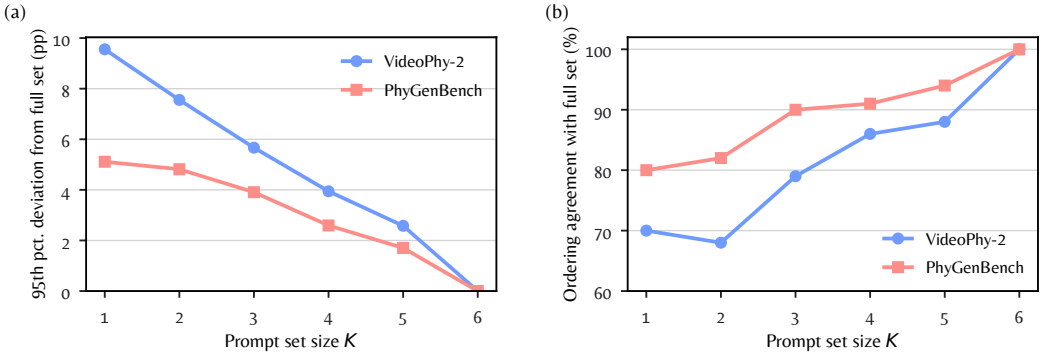


Figure S4: Finite-pool diagnostic over all subsets of size K that include the original prompt. (a) 95th percentile of the deviation of model scores from the full six-prompt set, in percentage points. (b) Mean agreement of pairwise model orderings with the full set. At $K = 5$ the agreement is 88.0% on VideoPhy-2 and 94.0% on PhyGenBench, and the 95th-percentile deviation is 2.6 and 1.7 percentage points.

F INDEPENDENTLY CONSTRUCTED PROMPT SET

Construction. The second prompt set uses the rewriting model and instruction of Appendix A in new generation contexts that never contain the first set. Candidates pass the same automatic filters, exact duplicates of the original prompt or of any first-set rewrite are removed, and the three language-model judges must all accept a candidate as equivalent. Of 1,008 candidates, 479 pass the filters and 469 are accepted by all judges. Five are selected per scenario, preferring candidates whose token edit distance to the nearest prompt of the first set is largest; all 30 scenarios reach five. The selected rewrites are on average 3.6 tokens from the nearest first-set prompt, whereas first-set rewrites are 2.7 tokens from their original. The second set is thus generated independently of the first; every candidate is accepted by all three judges before it is eligible, and selection prefers candidates far from the first set to avoid near-copies of it.

Generation and scoring. CogVideoX-5B and Wan2.1-14B generate one video per new rewrite with the settings and seeds of the main study (300 videos), and both evaluators score them as in the main study. All videos are retained, including five near-uniform CogVideoX-5B videos of one PhyGenBench scenario.

Results. Table S10 compares the five first-set rewrites (A) with the five second-set rewrites (B) of each scenario. The random-split reference pools the ten scores of each scenario, splits them into two groups of five at random, and averages the resulting set difference over scenarios (10,000 splits); p is the fraction of splits with a difference at least as large as the observed one. The one combination with a larger difference is Wan2.1-14B on PhyGenBench. Rescoring its set A with the hosted Stages 2 and 3 under the same setup as set B (Stage 1 is deterministic and was reused) changes 8 of the 90 final scores, none of them for the 15 original-prompt videos, and leaves the set difference at 10.2% [5.8, 15.1] against a random-split reference of 6.9 ($p = 0.04$), so it does not come from the time between the two scorings; Table S10 uses the set-A scores of the main study. Benchmark scores change little between the sets: from 4.31 to 4.24 (CogVideoX-5B) and from 3.99 to 3.97 (Wan2.1-14B) in mean PC score, from 37.3% to 40.0% and from 41.3% to 40.0% in VideoPhy-2 joint pass rate, and from 1.45 to 1.52 and from 1.47 to 1.40 in mean PhyGenBench score. The two models are close on both benchmarks: the CogVideoX-5B minus Wan2.1-14B difference in joint pass rate is -4.0 points [-10.7, 1.3] on set A and 0.0 [-8.0, 8.0] on set B, and in PhyGenBench score -0.01 [-0.33, 0.31] and 0.12 [-0.35, 0.55], so neither set resolves their order. Restricting the analysis to the 24 scenarios in which every second-set rewrite is at least two tokens from the first set leaves the scenario-level differences within 1.5 percentage points of those in Table S10.

Calibration. The uncertainty estimated from the first set predicts the second. For each scenario, model, and metric, the predicted standard deviation of the difference between the two five-prompt means is $\sqrt{2s^2/5}$, where s^2 is the within-scenario sample variance of the five first-set rewrites, pooled over the 15 scenarios of the model and metric. The reference values are obtained by splitting the ten rewrites of each scenario into two groups of five at random (10,000 splits). At the benchmark level, all six model–metric differences lie within one predicted standard deviation ($|z|$ from 0.16 to 0.94). At the scenario level, pooled over the six model–metric combinations, 66% and 88% of scenarios fall within the 68% and 95% prediction intervals, against 72% and 93% for random splits, and the mean absolute difference is 0.98 times its prediction against 0.81 for random splits; the differences larger than predicted come from PhyGenBench.

Table S10: First (A) versus independently constructed second (B) prompt set. *Changed*: fraction of rewrites whose result differs from that of the original prompt. *Single*: mean absolute difference between one prompt of A and one of B, over all 25 pairs. *Set*: absolute difference between the five-prompt means of A and B, with a 95% scenario bootstrap interval. *Random split*: the same difference expected when the ten prompts of a scenario are split into two groups at random. Differences are averaged over the 15 scenarios of a benchmark, in percent of the score range.

		Changed (%)		Difference between sets (% of range)			
Metric	Model	A	B	Single	Set	Random split	p
VideoPhy-2 PC	CogVideoX-5B	53.3	36.0	11.1	7.0 [3.7, 11.7]	5.8	0.23
	Wan2.1-14B	26.7	26.7	9.5	3.7 [2.0, 5.3]	5.8	>0.99
VideoPhy-2 joint	CogVideoX-5B	17.3	17.3	12.3	10.7 [4.0, 18.7]	8.2	0.28
	Wan2.1-14B	20.0	24.0	17.9	9.3 [4.0, 16.0]	11.4	0.90
PhyGenBench	CogVideoX-5B	37.3	42.7	16.1	9.3 [4.9, 14.2]	7.8	0.24
	Wan2.1-14B	24.0	25.3	12.7	10.2 [6.2, 14.7]	6.5	0.02

G Source of the Sensitivity

This section collects the evidence behind Section 4.4: the repeatability of the evaluators, the stages of the PhyGenBench evaluator, a fully repeatable local replacement for its hosted stages, and the human study.

G.1 Evaluator Repeatability

The repeat study of Section 4.4 scores every video five times with the released evaluator settings: the VideoPhy-2 checkpoint `videophy_2_auto` with greedy decoding, and for PhyGenBench the same procedure as in the main study (VQAScore in Stage 1; GPT-4o `gpt-4o-2024-05-13` in Stages 2 and 3, with the compact Stage-2 prompt for the two scenarios that use it). Eight PhyGenBench videos had at least one failed hosted call across their 25 stage evaluations and are excluded from the repeat statistics; no VideoPhy-2 evaluation failed. Intervals are 95% scenario-cluster bootstrap intervals over 10,000 draws.

Table S11: Evaluator repeatability versus the effect of rewriting. Every video is scored five times with identical inputs. *Unstable videos* is the fraction of videos whose score is not the same in all five evaluations (450 videos on VideoPhy-2; 127 on PhyGenBench after excluding 8 with a failed hosted call). The last two columns give the mean absolute score change, in percent of the score range, on the 40 score-changed pairs per benchmark that were sampled for the repeat study, independently of the human study (Appendix G.4). The two columns report the change between two evaluations of the same video, and between the original and the rewrite video at the same seed.

			Mean score change (%)	
Evaluator	Component	Unstable videos (%)	Same video, re-scored	Original vs. rewrite
VideoPhy-2	PC	0.0	0.0	27.6
	SA	0.0	0.0	6.4
PhyGenBench	Stage 1 (VQAScore)	0.0	0.0	22.1
	Stage 2 (GPT-4o)	17.3	2.7	19.2
	Stage 3 (GPT-4o)	24.4	8.2	40.4
	Final score	18.9	4.2	35.0

Table S11 summarizes the comparison. Over five repeats, the PhyGenBench final score changed for 18.9% [11.7, 25.9] of videos and the diagnostic decision for 11.8% [5.0, 19.2]; the Stage-3 score changed for 24.4% [14.8, 34.7] and the Stage-2 score for 17.3% [8.4, 27.4]. The mean absolute difference between repeats is 3.0% [1.8, 4.2] of the score range for the final score and 6.5% [3.8, 9.4] for Stage 3. The comparison of the repeat variation with the rewrite effect uses 40 score-changed pairs per benchmark that were sampled for the repeat study; this sample was drawn independently

of the human study (Appendix G.4), and the two share 27 of their 80 pairs. On these 40 pairs per benchmark, the inverse-probability-weighted repeat and rewrite mean absolute differences are 4.2% [2.1, 6.3] against 35.0% [33.3, 37.2] for the final score and 8.2% [3.4, 13.6] against 40.4% [22.9, 56.7] for Stage 3. On VideoPhy-2, all 2,250 PC and SA evaluations reproduced the first repeat exactly, so the corresponding repeat statistics are zero; a degenerate bootstrap interval does not bound unobserved disagreement.

Scoring all 450 VideoPhy-2 videos in the environment of the repeat study reproduces the main results (Appendix D.3). All analyses in the main text use the main-study scores.

G.2 PhyGenBench Evaluator Stages and Route

Table S12 reports, for each stage of the PhyGenBench evaluator, the variance of its score across the six videos of a scenario, averaged over scenarios. Table S13 repeats the main measurements on the 13 scenarios evaluated with the released Stage-2 procedure.

Table S12: Within-scenario variance of PhyGenBench stage scores (all on the 0–3 scale). The video-level stage varies most for every model.

Model	Stage 1 (single frame)	Stage 2 (multi-frame)	Stage 3 (video)	Final
CogVideoX-5B	0.250	0.441	**0.535**	0.156
SDXL+SVD-XT	0.228	0.370	**0.596**	0.180
Cosmos-Predict1-7B	0.156	0.170	**0.670**	0.254
Wan2.1-14B	0.202	0.259	**0.496**	0.181
HunyuanVideo-13B	0.128	0.148	**0.604**	0.231

Table S13: PhyGenBench results on the 13 scenarios evaluated with the released Stage-2 procedure (65 rewrites per model). Mean $|\delta|$ and Set range are percentages of the score range; Flip uses the diagnostic threshold.

Model	Mean $\|\delta\|$	Set range	Changed	Flip	Stage 2 var.	Stage 3 var.
CogVideoX-5B	13.3%	25.6%	40.0%	27.7%	0.509	0.532
SDXL+SVD-XT	13.3%	33.3%	40.0%	21.5%	0.410	0.652
Cosmos-Predict1-7B	16.4%	35.9%	47.7%	30.8%	0.197	0.744
Wan2.1-14B	10.3%	23.1%	24.6%	21.5%	0.282	0.442
HunyuanVideo-13B	13.8%	33.3%	38.5%	23.1%	0.171	0.679
All models	13.4%	30.3%	38.2%	24.9%	–	–

G.3 A Local Evaluator for the Hosted PhyGenBench Stages

To separate prompt sensitivity from the variation of hosted model calls, we re-evaluate all 450 PhyGenBench videos with Stages 2 and 3 judged by Qwen3.6-27B instead of GPT-4o. The model runs locally in BF16 with greedy decoding on one H100, without its extended reasoning mode, and receives exactly the same frames, questions, scenario-specific rubrics, and Stage-2 prompt variant as GPT-4o for each video; the instructions to reason step by step and to justify the choice are kept. Stage 1 is taken from the main study for both evaluators, and the final score is computed as before. Five of the 3,688 requests exceeded the output limit and are treated as missing rather than as scores.

Table S14 compares the two evaluators. Under Qwen, rewriting changes the Stage-2 score in 23.1% and the Stage-3 score in 32.5% of the comparisons, so the video-level stage is again the more sensitive one. Of the 360 comparisons, 58 change under both evaluators (53 in the same direction), 55 only under Qwen, 72 only under GPT-4o, and 175 under neither. Across videos, the two evaluators assign the same final score in 66.7% of cases (75.3% in Stage 2 and 59.2% in Stage 3); this agreement describes how similar the evaluators are and is not a measure of correctness. Exchanging the model also exchanges the deployment, so the absence of repeat variation is consistent with the variation of GPT-4o originating in the hosted service but does not isolate its cause.

Table S14: PhyGenBench with hosted GPT-4o or local Qwen3.6-27B in Stages 2 and 3. Repeat variation uses five evaluations of the 126 videos scored completely by both evaluators. Prompt sensitivity uses the 360 original-to-rewrite comparisons that are valid for both and exclude the flagged prompt pairs.

	GPT-4o	Qwen3.6-27B
Videos whose final score changes across five evaluations	19.0%	0.0%
Videos whose Stage-3 score changes across five evaluations	24.6%	0.0%
Rewrites that change the final score	36.1%	31.4%
Rewrites that flip the diagnostic decision	23.1%	18.9%

G.4 Human Video Evaluation

Items. The questionnaire contains 100 pairs of videos. Eighty are original/rewrite pairs whose physical score in the main study differs (VideoPhy-2 PC or PhyGenBench final score), 40 per benchmark, drawn by random sampling without replacement within each benchmark and two model groups (CogVideoX-5B and SDXL+SVD-XT; the other three models). Pairs that involve one of the four flagged rewrites of Appendix A, for any model, or contain a near-uniform video (Appendix H.2) are not eligible. Ten further pairs, five per benchmark, are drawn in the same way from the original/rewrite pairs whose score does not differ. Six are constructed control pairs: short animations of a bouncing ball, an undriven pendulum, and a collision of equal masses, in which one video follows the physical trajectory and the other injects energy or violates the conservation of momentum, with camera, scene, and initial conditions matched; an author confirmed each difference visually. Four pairs show the same video twice. All pairs were fixed before any rating was collected.

Task. For each pair, raters see the two videos side by side together with the original scenario description, and answer two questions: whether the videos realize the described scenario (both, only one, neither, or cannot judge), and which video is physically more plausible (the same, the first, the second, or cannot judge). "The same" means no apparent difference in physical plausibility; it requires neither identical trajectories nor that both videos are correct. Raters do not see the model, the prompts used for generation, which video came from the rewrite, or any score. The order of the pairs and the side of each video are randomized per rater.

Raters and aggregation. Four raters judged all 100 pairs. By their own declaration, none of them is an author or had seen the questionnaire, its pilot, or the control pairs before. A pair is assigned a judgment if at least three of the four raters give the same answer; answers of "cannot judge" remain in the denominator, and pairs without such a majority are reported as such. The bottom bar of Figure 6 uses the 32 pairs in which a majority judged both videos to show the described event.

Additional results. The share of score-changed pairs judged equally plausible is 51% with a 95% Wilson interval of [40, 62], and [38, 67] when scenarios are resampled within each benchmark with population weights. Of the 320 individual ratings of these pairs, 183 judge the two videos the same, 50 prefer the video with the higher score, 35 the other video, and 52 cannot judge. Leaving out one rater at a time, with a majority then requiring two of three raters, yields 43 to 50 equally plausible pairs out of 80 and 4 to 9 pairs with the evaluator's preference. Of the ten pairs with unchanged scores, seven are judged the same, one different, and two have no majority. On the six control pairs, 23 of 24 ratings identify the incorrect video, and on the identical pairs, 15 of 16 ratings are same and one cannot judge. The controls are animations with salient violations and are easier to judge than the generated videos. Raters agree with each other on 59% of pairwise comparisons for scenario realization (Krippendorff's $\alpha = 0.39$) and 56% for physical plausibility ($\alpha = 0.28$), which reflects how difficult physical plausibility is to judge in generated videos that often do not show the requested event.

By size and persistence of the score difference. The lack of support for the evaluator's preference is not limited to small or unstable score differences. Of the 80 pairs, 72 differ by one point of the raw score and 8 by two or more; the majority judgments are 39 same, 2 for the higher-scored video, 3 for the other video, and 28 without a majority for the one-point pairs, and 2, 0, 0, and 6 for the

larger differences. Persistence is judged from a second scoring of the same videos by the same evaluator: the rescored VideoPhy-2 scores of Appendix D.3 classifies all 40 VideoPhy-2 pairs, of which 37 keep the sign of their score difference, and the GPT-4o repeats of Appendix G.1 cover 14 of the 40 PhyGenBench pairs, of which 9 keep it. The 46 persistent pairs receive 22 same, 1 for the higher-scored video, 1 for the other video, and 22 without a majority; the 8 non-persistent pairs 4, 0, 0, and 4.

H Prompts versus Generation Seeds

The crossed experiment of Section 4.5 uses five VideoPhy-2 scenarios drawn uniformly at random from the 13 scenarios without a flagged prompt, the six prompts of each scenario, and seeds 101–105 shared by all prompts, scenarios, and both models (300 videos, all generated and scored; one Wan2.1-14B generation failed for a technical reason and was regenerated with the same seed). Scenarios are weighted equally, and intervals resample whole scenario tables (10,000 draws), so they describe the variation over scenarios; the tests of Appendix H.1 treat the seeds as random blocks. The allocation comparison enumerates all possible draws of each strategy from the 6×5 table of a scenario, with the mean of that table as the target.

Table S15 summarizes the experiment. The seed effect is slightly larger than the prompt effect in both models, but the paired intervals of the difference include zero (-0.9 $[-4.4, 2.3]$ for CogVideoX-5B and -1.3 $[-3.0, 0.0]$ for Wan2.1-14B), and individual differences are almost always a single point on the 1–5 scale for both factors. After seed averaging, the score of at least one rewrite differs from that of the original in five of five and three of five scenarios, and no rewrite is better or worse than the original in all five seeds. In the allocation comparison, five prompts with five distinct seeds estimate the PC score with an RMSE of 3.8% for CogVideoX-5B against 5.3% for one prompt with five seeds, a paired difference of -1.5 $[-2.8, -0.2]$; for Wan2.1-14B the two allocations give 1.97% and 2.03%, a difference of -0.06 $[-0.37, 0.19]$, which is not resolved. Spending all five generations on prompts or all on seeds gives similar errors for CogVideoX-5B (5.6% and 5.3%), and for Wan2.1-14B the difference of 1.2 $[0.0, 2.6]$ between five prompts at one seed and one prompt at five seeds is not resolved either. On the official joint decision, the pass outcome is constant within most scenarios (passes out of 30 videos per scenario: 0, 0, 0, 29, and 12 for CogVideoX-5B and 0, 0, 1, 30, and 30 for Wan2.1-14B), and no allocation difference is resolved for either model; the joint variation of CogVideoX-5B comes almost entirely from near-uniform videos (Appendix H.2). The mean absolute change of the joint pass indicator between prompts at the same seed, between seeds at the same prompt, and between the original and a rewrite after seed averaging is 9.9, 8.0, and 7.2 percentage points for CogVideoX-5B and 1.3, 1.3, and 0.8 for Wan2.1-14B.

Table S15: Prompts versus generation seeds on VideoPhy-2 (five scenarios, six prompts, five seeds per model). Left: mean absolute change of the PC score, in percent of the score range, between two prompts at the same seed, between two seeds at the same prompt, and between the original prompt and a rewrite after averaging over the five seeds; brackets are 95% bootstrap intervals over scenarios. Right: root-mean-square error of four ways of spending the generations of a scenario when estimating its mean score over all prompts and seeds; each five-video allocation uses five generations, and "5 prompts, 5 seeds" pairs five distinct prompts with five distinct seeds (five videos, not 25).

	Mean absolute change of PC (%)			Estimation error (RMSE, %)			
Model	Prompts, same seed	Seeds, same prompt	After seed averaging	1 video	5 prompts, 1 seed	1 prompt, 5 seeds	5 prompts, 5 seeds
CogVideoX-5B	7.5 [5.6, 9.7]	8.3 [5.7, 11.3]	5.6 [4.0, 7.2]	10.9	5.6	5.3	**3.8**
Wan2.1-14B	3.2 [0.7, 5.7]	4.5 [0.7, 8.3]	2.4 [0.4, 4.4]	5.5	3.2	2.0	**2.0**

Table S16 gives the per-scenario effects. Which factor is larger depends on the scenario, and for Wan2.1-14B two scenarios receive the same PC score for all 30 videos. A two-way decomposition of each table assigns 25% of the sum of squares to prompts, 25% to seeds, and 50% to the residual for CogVideoX-5B, and 14%, 30%, and 55% for Wan2.1-14B (scenarios with constant scores excluded); with one video per cell, the residual combines the prompt–seed interaction with any remaining generation and evaluation variation.

Examples. Figure S5 shows, for each model, three videos generated from the same prompt with different seeds. For CogVideoX-5B, all three show a paint roller covering a textured wall and receive

Table S16: Mean absolute change of the PC score (percent of the score range) between prompts at a fixed seed and between seeds at a fixed prompt, by scenario.

	CogVideoX-5B		Wan2.1-14B	
Scenario	Prompts	Seeds	Prompts	Seeds
Climber falling on a rope	11.0	6.7	6.0	8.3
Javelin landing in mud	9.3	10.8	6.7	10.8
Toothpaste squeezed from a tube	6.0	5.8	3.3	3.3
Paint roller on a wall	6.0	13.3	0.0	0.0
Syrup poured on a pancake	5.0	5.0	0.0	0.0
Mean	7.5	8.3	3.2	4.5

the maximum SA score, yet the PC score is 5 for one seed and 4 for the other two; the highest PC score goes to the video in which the roller leaves the frame early and the wall is static for most of the clip. For Wan2.1-14B, all three videos show toothpaste being squeezed onto a toothbrush in a continuous stream, yet one seed receives PC and SA scores of 4 and passes, while the other two receive 3 and fail. Neither case exhibits an apparent difference in the physical plausibility of the depicted event that would account for the different scores, which suggests that part of the seed effect, like part of the prompt effect, reflects the evaluator's response to the appearance of a video. These examples were selected for illustration among prompts whose PC score differs across seeds (the largest difference observed in this experiment is one point) after excluding the near-uniform videos of Appendix H.2; the human study of Section 4.4 addresses the same question systematically for the prompt effect.

H.1 Seed-Aware Tests of the Prompt Effect

For CogVideoX-5B the prompt changes the expected PC score beyond seed variation, for Wan2.1-14B it does not, and for neither model does the original prompt differ systematically from its rewrites (Table S17). The crossed experiment gives, per model and scenario, a 6×5 table of prompts and seeds with one video per cell, so the prompt–seed interaction is part of the error term of every test. The prompt main effect is tested by permuting the prompt labels independently within each seed (20,000 permutations), with the variance of the six prompt means as the statistic; the null hypothesis is no prompt main effect with arbitrary seed effects. The per-scenario statistics are summed for a stratified test over the five scenarios, and a Friedman test with seeds as blocks gives the same picture. The original prompt is compared with its rewrites by the seed-paired difference between its score and the mean of its five rewrites, with a seed bootstrap interval (10,000 draws), and by exact label tests in which each of the six prompts in turn plays the role of the original. Variance components for prompt, seed, and residual come from a two-way decomposition with scenario means removed, with two-level bootstrap intervals (5,000 draws).

Table S17: Seed-aware tests of the prompt effect on the PC score. Stratified p is the within-seed permutation test of the prompt main effect pooled over the five scenarios; Scenarios counts scenarios with $p < 0.05$ on their own. Standard deviations are in percent of the score range with 95% bootstrap intervals; the residual includes the prompt–seed interaction. The last column is the seed-averaged difference between the original prompt and the mean of its rewrites, with a 95% seed bootstrap interval.

Model	Stratified p	Scenarios	Prompt SD	Seed SD	Residual SD	Original − rewrites
CogVideoX-5B	0.004	1 of 5	4.5 [2.1, 8.2]	5.4 [2.6, 8.1]	9.1 [4.9, 9.7]	−1.6 [−3.4, 0.0]
Wan2.1-14B	0.53	0 of 5	0.9 [0.0, 4.9]	4.1 [0.0, 6.7]	6.3 [3.2, 6.6]	−0.8 [−2.8, 1.8]

For CogVideoX-5B, the stratified permutation test rejects the absence of a prompt effect ($p = 0.004$); the climber scenario has $p < 0.05$ on its own, and the effect remains when the near-uniform videos of Appendix H.2 are removed ($p < 0.0001$). The prompt standard deviation is 4.5% of the score range, 15% of the variance, against 5.4% for the seed (22%) and 9.1% for the residual (62%). For Wan2.1-14B, the stratified test gives $p = 0.53$, no scenario has $p < 0.05$, two scenarios are constant

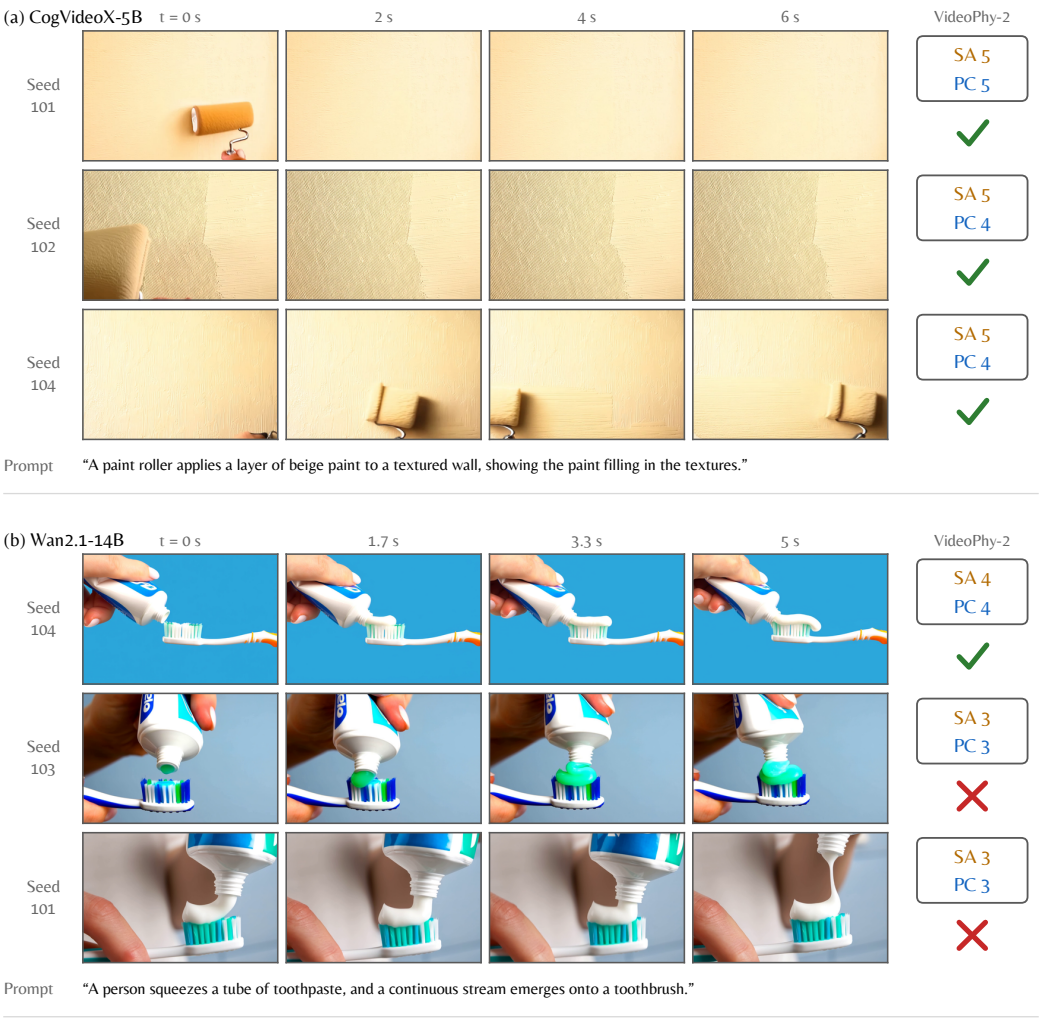


Figure S5: Same prompt, different generation seeds. (a) CogVideoX-5B: three seeds yield comparable videos of a paint roller on a textured wall with identical SA scores, but the PC score differs. (b) Wan2.1-14B: three seeds yield comparable videos of toothpaste squeezed onto a toothbrush, but the scores differ and the official decision changes from pass to fail. Frames are sampled at equal intervals and cropped identically; the prompt is one of the validated rewrites of the scenario.

over all 30 videos, and the prompt standard deviation is 0.9% of the range, so its differences between seed-averaged scores are within seed variation. The pooled seed-paired difference between the original prompt and its rewrites is -1.6% $[-3.4, 0.0]$ of the range for CogVideoX-5B and -0.8% $[-2.8, 1.8]$ for Wan2.1-14B, and the original is no more different from the rewrites than any rewrite is from the others (label tests, p from 0.35 to 0.82). On the joint decision, which is constant in most scenarios, the one prompt effect (CogVideoX-5B, syrup scenario, $p = 0.005$) is the near-uniform failure: all 17 flagged videos fail, 12 of the 13 unflagged videos pass, and the one rewrite that never produced a near-uniform video passes at all five seeds.

H.2 Maximum Physical Scores on Near-Uniform Videos

Thirty-six of the 150 CogVideoX-5B videos of the seed experiment (24%) are nearly uniform in color and show no recognizable event: 18 in the toothpaste scenario, 17 in the syrup scenario, and one in the paint-roller scenario (Figure S6). They were flagged automatically during generation, before any scoring, as videos whose spatial standard deviation of pixel values, averaged over channels and frames, is below 2 on the 0–255 scale. None of the Wan2.1-14B videos is flagged. *All 36 receive*

the maximum PC score of 5 from the VideoPhy-2 evaluator, whose PC question does not contain the prompt, while their SA scores are 2 or 3, so none passes the joint decision. The same rule flags 15 of the 900 videos of the main study, again all from CogVideoX-5B (nine on VideoPhy-2 and six on PhyGenBench) and none from the other four models, so these outputs are a property of this model rather than of one inference environment.

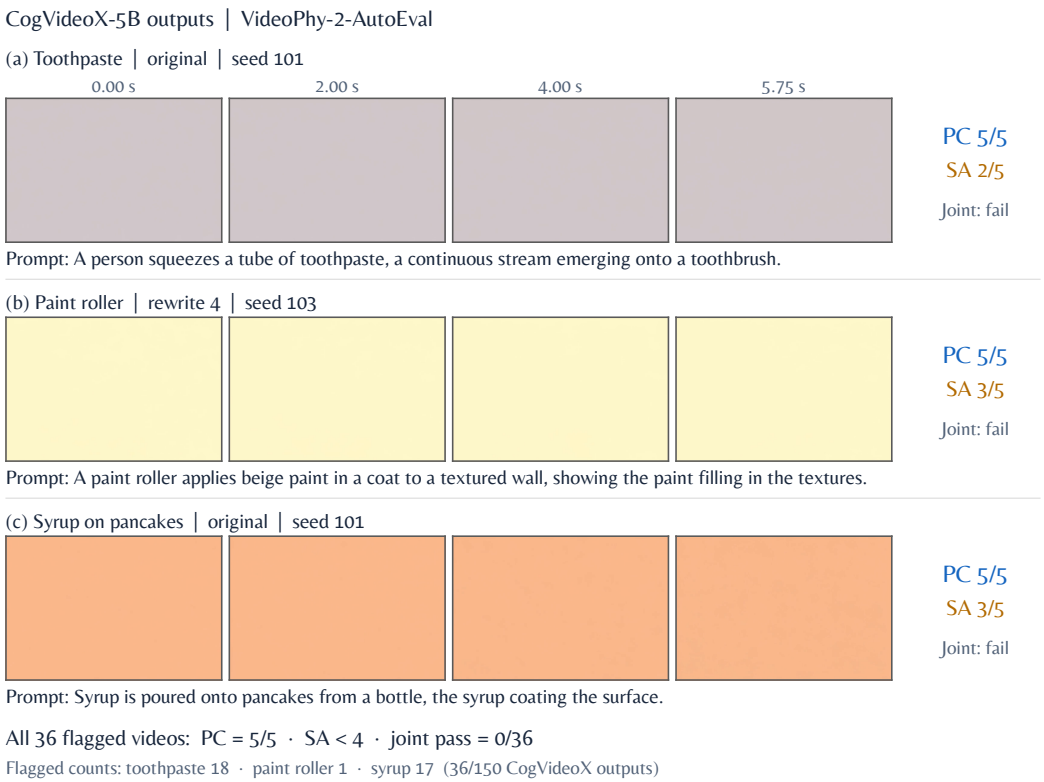


Figure S6: Near-uniform CogVideoX-5B videos receive the maximum PC score from the VideoPhy-2 evaluator. One flagged video is shown for each affected scenario, with unprocessed frames and the prompt used for generation. PC is 5 although no event is visible; SA is below 4, so the joint decision fails. The same holds for all 36 flagged videos.

The observation shows that a caption-free physical judgment can rate a video without the requested event, or any event, as fully plausible, presumably because nothing in it violates physics. This does not affect the benchmark's own conclusions: VideoPhy-2 reports the joint performance, which requires $\mathrm{SA} \geq 4$ in addition to $\mathrm{PC} \geq 4$, and this criterion rejects all 36 videos. It warrants caution when using the PC score in isolation, including our own analyses of PC changes, which is why we report the official joint decision alongside PC throughout.

Effect on the main study. The nine flagged VideoPhy-2 videos of the main study show the same pattern: all receive a PC score of 5 and an SA score of 2 or 3, and none passes the joint decision. On PhyGenBench, the six flagged videos belong to one scenario; their first two stages score 0, yet the video-level GPT-4o stage assigns four of them its maximum score of 3, and the final score is 0 or 1. Excluding every comparison that involves a flagged video changes the results of Section 4.2 by less than half a percentage point: the physical score changes in 35.6% of the remaining 360 VideoPhy-2 comparisons (35.5% of all 375) and the decision flips in 13.1% (13.3%); on PhyGenBench the corresponding values are 35.7% and 22.7% of 370 comparisons (35.7% and 22.4%).

Effect on the comparison of prompts and seeds. All videos are kept in the analyses of Section 4.5, since removing outputs by content would change the evaluated population. Table S18 repeats the comparison using only pairs in which neither video is flagged. Both effects remain, and they increase for CogVideoX-5B, because pairs of flagged videos always tie at the maximum score and thereby

hide variation. The order of the two effects reverses for CogVideoX-5B, which confirms that neither factor consistently dominates. The filtered design is unbalanced (only 7 of the 60 seed pairs of the toothpaste scenario remain), so these values complement rather than replace the full analysis.

Table S18: Prompt and seed effects with and without near-uniform videos: mean absolute change of the PC score in percent of the score range, with equal scenario weights. A pair is retained only if neither video is flagged.

Model	Varied factor	All pairs	Unflagged pairs	Retained pairs
CogVideoX-5B	Prompt	7.5	10.4	264/375
	Seed	8.3	9.0	197/300
Wan2.1-14B	Prompt	3.2	3.2	375/375
	Seed	4.5	4.5	300/300