PhysVidBench: Language-Grounded Evaluation of Physical Commonsense in Text-to-Video Models

Enes Sanli[1,2,*†] **Baris Sarper Tezcan**[3,*] **Erkut Erdem**[2,4] **Aykut Erdem**[1,2]
[1]Koç University, Istanbul, Turkey [2]KUIS AI Center, Istanbul, Turkey
[3]Purdue University, Indiana, USA [4]Hacettepe University, Ankara, Turkey
[*]Equal contribution.
[†]**Correspondence:** esanli25@ku.edu.tr

Abstract

Text-to-video (T2V) models now produce striking visuals, yet they routinely violate everyday physics; objects float, tools are misused, and causal sequences break down. Existing benchmarks mostly probe isolated physical laws and rely on vision-language models to score videos directly, which entangles perception and reasoning in one judgment and correlates poorly with humans. We introduce PhysVidBench, a benchmark of 383 base prompts, expanded to 766 prompts with enriched variants and 4,123 manually reviewed QA items, together with a language-grounded evaluation framework that takes a different route: instead of asking a VLM "Is this video physically correct?", we caption the video, then ask a language model to answer prompt-derived yes/no physics questions using only the captions. This split between seeing and reasoning makes each score traceable to the captions that justify it and aligns more closely with human judgment than direct VLM scoring (Pearson $r = 0.45$–0.69, compared with 0.21–0.47 for the strongest direct VLM evaluator). To test the framework, we carefully curate a set of human-validated prompts spanning seven physical dimensions, with a focus on tool use and affordances – areas largely absent from prior physics-focused benchmarks. Across 12 open and proprietary T2V systems, the best model reaches only 36.2% average accuracy, and no model consistently handles everyday physical reasoning. The same pipeline can also guide iterative error-guided prompt refinement, improving CogVideoX-2B from 21.6 to 32.7 and CogVideoX-5B from 17.8 to 29.7 without retraining.

1 Introduction

To separate an egg yolk from the white, one can squeeze an empty plastic bottle, place its mouth over the yolk, and release it so that suction lifts the yolk (Figure 1). This simple action requires commonsense about deformability, suction, spatial alignment, and temporal order. For T2V models, the challenge is not merely rendering a visually appealing scene, but realizing the physical interactions implied by language. Assessing whether they succeed is itself difficult. It requires checking whether the video evidence supports the prompt's causal and procedural expectations.

Recent T2V models produce increasingly realistic videos (NVIDIA et al., 2025a; Kong et al., 2025; Yang et al., 2025; Chen et al., 2024; Team Wan et al., 2025; Kuaishou Team, 2024; Germanidis, 2024; van den Oord and Roman, 2024) and are often discussed as possible world models for robotics, embodied AI, and simulation-based learning (He et al., 2026; Hu et al., 2026; Wang et al., 2024; Hu et al., 2023a; Liu et al., 2024b). Yet visual realism does not imply physical understanding. Generated videos may preserve object appearance while failing to capture affordances, causal effects, material behavior, or the order of actions. For physical requirements expressible as discrete, visually verifiable propositions, this creates an evaluation problem that is naturally language-centered: the expected behavior is described linguistically, while the evidence must be extracted from video.

Existing benchmarks provide important progress, but leave two gaps. First, many focus on isolated physical laws, motion plausibility, or predefined temporal changes rather than everyday, tool-mediated physical interactions (Meng et al., 2025; Bansal et al., 2026; Zheng et al., 2025). Second, direct VLM-based scoring can conflate perception, grounding, and reasoning, making it hard to inspect why a video was judged correct or incorrect.

We introduce *PhysVidBench*, a benchmark and language-driven evaluation framework for physical commonsense in generated videos. It contains 383 prompts adapted from PIQA (Bisk et al., 2020), covering everyday interactions such as tool use, container manipulation, material transfer, and

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 45311–45359
October 24-29, 2026 ©2026 Association for Computational Linguistics

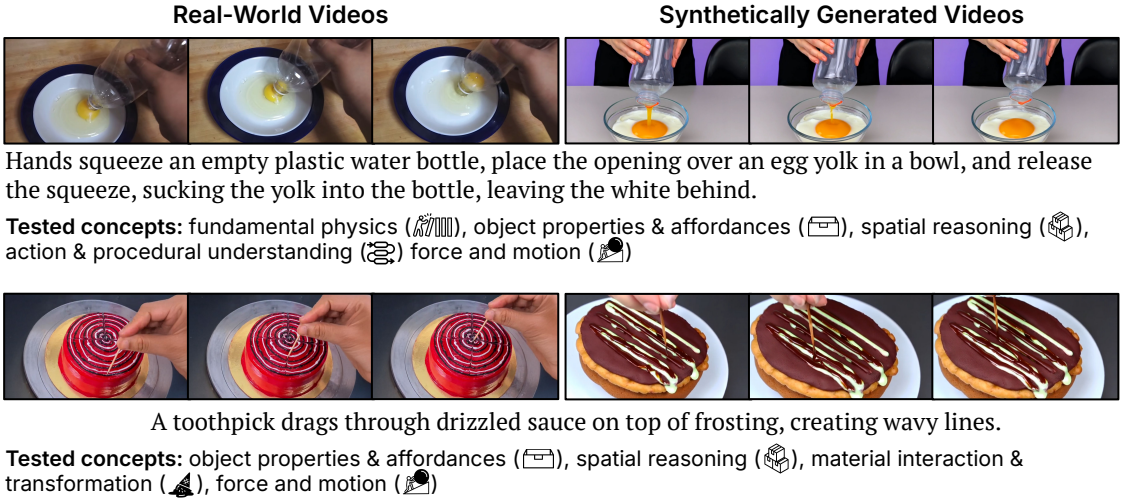


Figure 1: **Understanding physical commonsense in video generation.** Humans intuitively grasp how everyday interactions unfold: how objects respond to forces, how tools function, and how materials behave under manipulation. *PhysVidBench* tests whether T2V models share this commonsense, evaluating whether a generated video exhibits the physical consequences a prompt implies. Each row shows a real-world video of the correct behavior alongside a model's attempt at the same scenario. The generation looks plausible but misses the intended physical effect(s); the yolk is not drawn into the bottle, and the toothpick leaves no coherent pattern.

causal object effects. Each prompt is paired with a detail-enriched variant that makes the expected physical behavior more explicit, and the prompts are organized around physical reasoning dimensions (Figure 2), from force and motion and object affordance to material interaction, temporal progression, and procedural understanding. These dimensions are multi-label rather than mutually exclusive: a prompt may belong to multiple dimensions through its reviewed QA items (see App. F for an end-to-end example).

The core of our approach is a *caption-grounded question-answering evaluator*. Given a prompt, we derive yes/no questions about the expected physical behavior. A VLM captions the generated video, and an LLM answers using only this evidence. This decouples perception from reasoning: captions describe observable content, while the LLM checks whether they entail the required behavior. Unlike direct VLM judgments, the resulting decisions are traceable to explicit textual evidence, making evaluation more modular and interpretable.

We manually validate the generated questions and evaluate a diverse set of open- and closed-source T2V models. Across 12 systems, the best average accuracy is only 36.2%. Our results show that current models still struggle with everyday physical commonsense, especially for spatial relations, temporal dynamics, material interactions, and tool-mediated causal effects. We further show that caption-grounded QA aligns more closely with human judgments than direct VLM scoring and generic automatic metrics. On the same 120 videos and human annotations, our evaluator achieves Pearson correlations of 0.45–0.69, compared with 0.21–0.47 for the strongest direct VLM evaluator.

Our contributions are listed as follows:

- We introduce PhysVidBench, a PIQA-derived benchmark of 383 base prompts, 383 enriched variants, and 4,123 manually reviewed QA items that evaluates whether generated videos satisfy language-specified everyday physical commonsense.
- We formulate generated-video evaluation as caption-grounded QA, separating visual evidence extraction from language-based physical reasoning for physical requirements expressible as discrete, visually verifiable propositions.
- We benchmark multiple T2V models, both open and proprietary, and show persistent physical-reasoning failures, demonstrate stronger human alignment than direct VLM scoring, and show the same pipeline can guide training-free prompt refinement.

All 766 base and upsampled prompts, the 4,123 reviewed QA items, captioning, and judge prompts, scoring code, and documentation of the curation and evaluation procedures are available at https://cyberiada.github.io/PhysVidBench/.

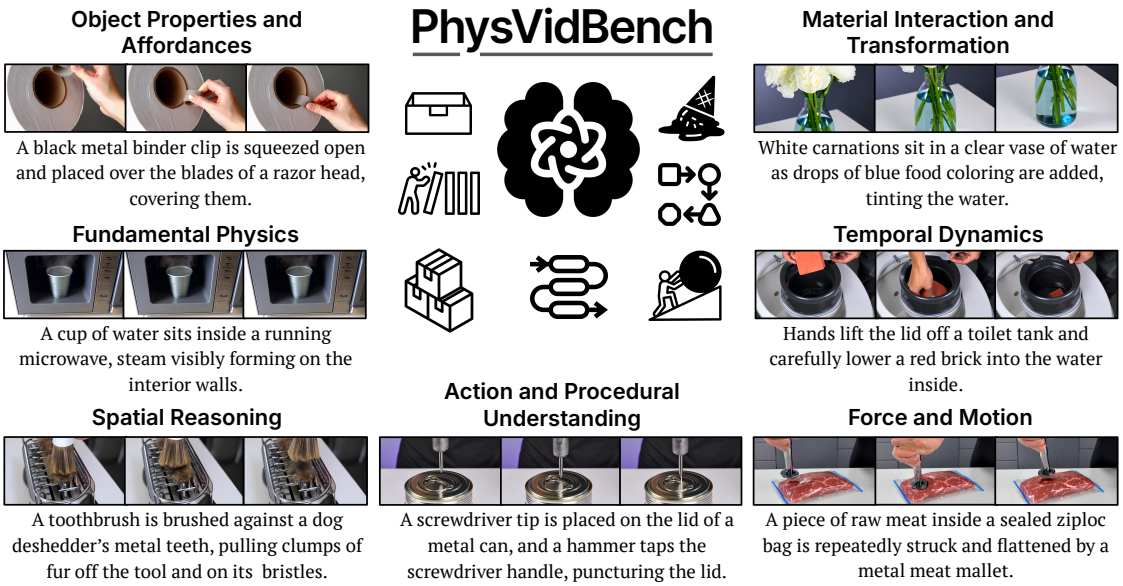


Figure 2: **Physical commonsense dimensions tested in PhysVidBench**, each illustrated with a video generated by Cosmos-14B. Prompts probe force and motion, object affordance, spatial containment, temporal progression, material interaction, and related requirements. The dimension shown for each example is illustrative rather than exclusive; prompts are annotated at the QA level and a single prompt typically contributes to multiple dimensions.

2 Related Work

Text-to-Video Generation and Its Evaluation. Text-to-video (T2V) models, such as CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2025), VideoCrafter2 (Chen et al., 2024), Cosmos (NVIDIA et al., 2025a), Wan (Team Wan et al., 2025), Sora (Brooks et al., 2024), and Veo (van den Oord and Roman, 2024), now generate increasingly realistic and temporally coherent videos. Yet realism and prompt alignment do not guarantee that a video satisfies the physical interaction described by the prompt. Early metrics such as CLIP Score (Hessel et al., 2021), FVD (Unterthiner et al., 2019), and IS (Salimans et al., 2016) capture coarse alignment or distributional quality, but are weak proxies for temporal consistency, motion plausibility, and physical validity. Recent benchmarks provide finer-grained evaluation (Huang et al., 2024b; Liu et al., 2024a, 2023), but remain limited when correctness hinges on tool use or physical effects.

Vision-Language Models as Evaluators. An emerging trend is to use VLMs as automatic evaluators for generated content (Huang et al., 2024b; Zheng et al., 2025), and on the dimensions they target (e.g. visual quality, temporal smoothness, or object presence) VLM scores correlate well with human ratings. That said, physical commonsense is a harder regime. The evaluator must decide whether the video supports a language-implied causal event, not merely whether the relevant objects appear, and prior analysis places physical reasoning closer to hallucination-sensitive evaluation than to standard VLM benchmarks (Chow et al., 2025). Folding perception, grounding, and reasoning into one model call then makes scores hard to interpret and prone to hallucination, superficial matching, or response collapse (Huang et al., 2024a). In App. B, we show such scoring can fail to produce meaningful judgments on physical prompts.

Prior work has decomposed generated-content evaluation into fine-grained questions or structured descriptions, including TIFA, DSG, Divide-Evaluate-Refine, LLMScore, T2VScore, and ETVA (Hu et al., 2023b; Cho et al., 2024; Singh and Zheng, 2023; Lu et al., 2023; Wu et al., 2024; Guan et al., 2025). We do not claim question decomposition itself as novel; our key distinction is the intermediate caption-evidence layer, which separates visual evidence extraction from physical reasoning. Related caption-then-reason pipelines have also been effective in long-range video question answering (Zhang et al., 2024).

Instead, we decompose the judgment where, following the LLM-as-judge paradigm (Zheng et al., 2023), a captioning step sits between perception and reasoning, converting each video into caption evidence over which an LLM answers physics-grounded yes/no questions. Since the judge reasons

Table 1: **Physical commonsense video benchmarks.** PhysVidBench emphasizes tool use and affordances, compares base and detail-enriched prompts, and evaluates generated videos through caption-grounded Yes/No QA.

Feature	PhyGenBench	VBench-2.0	VideoPhy-2	PhysVidBench (Ours)
# of Prompts	160	1013	3940	383
Tool & Affordance Focus	No	No	No	Yes
Human Alignment Study	Yes	Yes	Yes	Yes
Upsampled Prompt Comparison	Yes	No	Yes	Yes
Automatic Evaluation	Yes	Yes	Yes	Yes
Human Feedback Type	Rating	Pairwise	Rating (1-5)	Yes/No QA

over text rather than raw video, decisions remain traceable to explicit captions, reformulating T2V assessment as constrained caption-grounded QA and reducing reliance on opaque VLM judgments.

Physical Commonsense Benchmarks and Reasoning. Physical commonsense has been studied in both language and vision-language models. PIQA tests everyday physical reasoning through goal-solution pairs involving tools, objects, and affordances, while PhysBench (Chow et al., 2025) and Impossible Videos (Bai et al., 2025c) show that VLMs and video models still struggle with physical world knowledge and implausible events. Beyond understanding, several benchmarks probe physical commonsense in *generated* video. VBench-2.0 evaluates dimensions such as state change, geometry, and motion rationality; VideoPhy and VideoPhy-2 (Bansal et al., 2025, 2026) focus on action-centric plausibility; PhyGenBench (Meng et al., 2025) targets isolated laws such as buoyancy and friction; and Physics-IQ (Motamed et al., 2026) and Cosmos-Reason1 (NVIDIA et al., 2025b) evaluate broader physical and causal reasoning phenomena.

As summarized in Table 1, PhysVidBench departs from these benchmarks in three important ways. First, it focuses on everyday tool-mediated interactions and affordance reasoning, where physical correctness depends on how objects are used, transformed, and causally related. Second, it pairs each base prompt with a detail-enriched variant to study how explicit physical constraints affect generation, while using aggregate model performance to organize prompts into Medium, Hard, and Very Hard groups for analysis. Third, it casts evaluation as yes/no QA, treating physical commonsense not merely as a capability probed in video but as a structured, language-grounded evaluation problem. Please refer to App. A for additional details.

3 Benchmark

PhysVidBench targets everyday physical commonsense: whether objects are used according to their affordances, actions unfold in the correct order, and physical effects such as suction, cutting, folding, pouring, or material deformation are actually realized. This makes the benchmark both a video-generation task and a language-mediated evaluation problem. Prompts define the expected behavior, and generated videos are assessed through caption-grounded question answering.

We build PhysVidBench from PIQA (Bisk et al., 2020), a dataset of everyday goal-solution pairs designed to test physical commonsense, mining it for the subset whose solutions involve tool use and non-obvious object affordances rather than abstract physical laws or synthetic scenes (App. C). PIQA contains roughly 21K examples. Using the labeled train and validation splits yields more than 17K correct goal-solution pairs, from which we retain 383 final prompts. We organize these scenarios along a seven-dimensional physical commonsense ontology (Table 2) spanning fundamental physics, object properties and affordances, spatial reasoning, temporal dynamics, action and procedural understanding, material interaction and transformation, and force and motion. The ontology serves two roles: it guides prompt selection and structures the evaluation pipeline, where questions and dimension-specific captions target the same physical reasoning dimensions.

3.1 Prompt Construction

Each prompt describes a short, concrete action that can be directly visualized and reasoned about. Construction proceeds in three stages: selecting and filtering physically grounded goal-solution pairs, generating base video prompts, and enriching them to surface the underlying physical properties and causal structure (Figure 3).

Table 2: **Physical commonsense ontology and prompt coverage in PhysVidBench.** The seven dimensions guide question generation and dimension-specific captioning. Coverage counts are non-exclusive: a prompt is counted toward every dimension for which it contains at least one manually reviewed QA item.

Commonsense Dimension	Definition	# Prompts Covered
Fundamental Physics	Core physical principles such as energy conservation, causality, equilibrium, and state transitions in physical systems.	349
Object Properties & Affordances	Inherent object attributes including material composition, rigidity, softness, and functional affordances like containment or support.	371
Spatial Reasoning	Spatial relations including position, geometry, occlusion, orientation, and fit between objects in a scene.	363
Temporal Dynamics	Timing, sequencing, and the causal structure of events over time, such as ordering, delays, and waiting.	209
Action & Procedural Understanding	Structured, goal-driven sequences of actions involving procedural steps toward task completion.	335
Material Interaction & Transformation	Responses of materials to external forces or processes including transformations like melting, freezing, breaking, or chemical change.	288
Force and Motion	Physical interactions governed by forces, including motion, pushing, pulling, lifting, and properties like inertia.	370

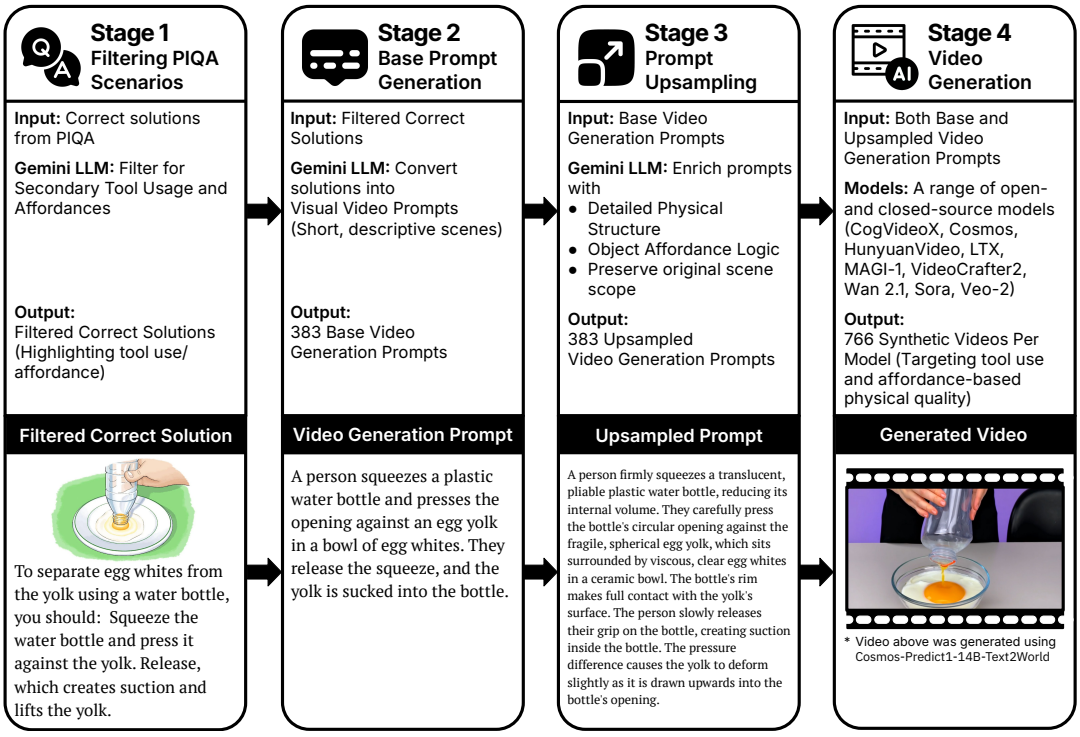


Figure 3: **Our prompt construction pipeline.** PIQA goal-solution pairs are filtered for tool use and non-obvious affordances, converted into video prompts, and expanded into detail-enriched variants before T2V generation.

Stage 1: Filtering PIQA Scenarios. Using Gemini-2.5-Pro-Preview-05-06 (Google DeepMind, 2025b) (hereafter referred to as Gemini 2.5 Pro), we retain PIQA pairs whose labeled-correct solution involves secondary tool use or non-obvious object affordances, such as using a credit card to scrape ice or a paperclip to eject a SIM tray. This selects for interactions where physical correctness depends not on recognizing objects, but on understanding how their properties enable a particular action.

Stage 2: Base Prompt Generation. Each retained pair is converted into a concise video prompt using Gemini 2.5 Pro. Prompts must describe realistic, self-contained, directly observable actions such as cutting, lifting, folding, or pouring; non-visual states, abstract intentions, and scenarios that cannot be shown in a short video are excluded. This yields 383 base prompts spanning more than 1,200 distinct objects and over 600 manipulation actions.[1]

[1]Object and action counts are obtained by parsing prompts with Gemini 2.5 Pro; see App. A.2.

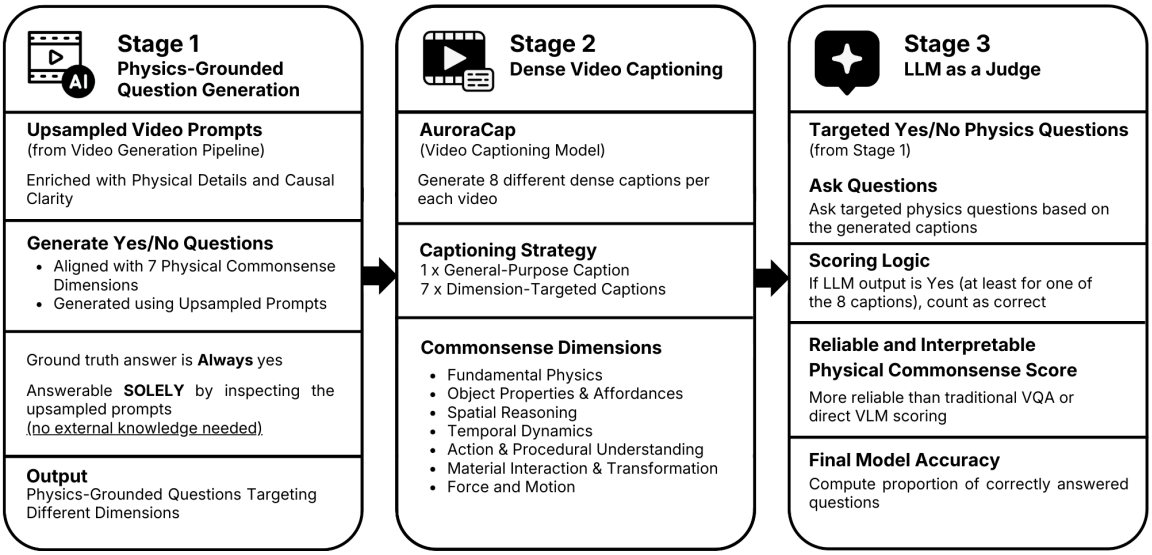


Figure 4: **Caption-grounded QA evaluation.** PhysVidBench converts each prompt into physical yes/no questions, captions each generated video with general and dimension-specific descriptions, and uses an LLM judge to answer using only this textual evidence.

Stage 3: Prompt Upsampling. We pair each base prompt with a detail-enriched variant, following prompt upsampling methods (Xue et al., 2025). Enrichment makes *implied* physical properties and causal structure explicit, for instance naming the suction or deformation an action already entails, while preserving the original scene scope; the prompts include instructions to avoid new objects, events, or speculative consequences. The final set contains 383 base and 383 enriched prompts (766 total).[2]

The prompts were manually validated by four authors, and each item was independently reviewed by two authors. Further curation details are provided in App. C.

3.2 Evaluation Pipeline

PhysVidBench evaluates generated videos with the caption-grounded QA pipeline in Figure 4. Rather than scoring a video with a single VLM call, we separate visual grounding from textual reasoning. A video is first converted into caption evidence, and an LLM then answers physics-grounded yes/no questions using only that evidence (App. D). This turns generated-video evaluation into a constrained language-reasoning problem whose decisions can be traced back to explicit captions.

Stage 1: Physics-Grounded Question Generation. For each enriched prompt, we generate targeted yes/no questions about the expected physical behavior. Each question is posed so that the correct answer is always "yes". It asks whether a required event, relation, affordance, or transformation occurs. This is a deliberate design choice. A physical interaction can fail in unboundedly many ways, so enumerating "no" cases yields ill-defined targets, whereas asking whether the intended effect is present gives each question a single, well-grounded positive answer.

If a prompt requires an object to fall, we ask whether the falling is supported by the video, rather than enumerating the many ways a generation might violate physics. Because the questions are derived from the enriched prompt, both base and enriched videos are evaluated against the same physical expectations, making their scores directly comparable. The process yields 4,123 questions, averaging 11 per video.[3]

Stage 2: Dense Video Captioning. Each generated video is captioned with AuroraCap (Chai et al., 2025). For every video, we produce one general-purpose caption and seven dimension-specific captions aligned with the ontology, each prompting the captioner to describe evidence relevant to one dimension, such as object affordances, temporal progression, material behavior, or force-based interactions. Together, these eight captions form the

[2]All prompts are manually reviewed to remove hallucinated, unsafe, inappropriate, or visually underspecified cases before T2V generation.

[3]All questions are manually reviewed for clarity, answerability, and visual grounding; about 5% are discarded because they refer to non-visual details or to information not shared by the base and enriched prompts.

Table 3: **Performance by physical commonsense dimension.** Accuracy is reported on upsampled prompts; parentheses show the change when switching to base prompts. Dimensions follow Table 2: **AU** = Action & Procedural Understanding, **FM** = Force and Motion, **FP** = Fundamental Physics, **MT** = Material Interaction & Transformation, **OP** = Object Properties & Affordances, **SR** = Spatial Reasoning, and **TD** = Temporal Dynamics. Bold and underline mark the best and second-best methods within each column.

Model	AU	FM	FP	MT	OP	SR	TD	Average
LTX-Video	20.7 (-4.2)	19.9 (-3.5)	18.6 (-2.5)	16.9 (-1.6)	25.0 (-4.7)	17.3 (-2.3)	13.7 (-4.6)	21.0 (-3.0)
VideoCrafter2	17.7 (-1.4)	19.4 (-2.9)	23.3 (-3.3)	18.8 (-2.9)	26.6 (-2.1)	15.8 (-0.2)	12.6 (-2.7)	22.1 (-1.7)
CogVideoX (2B)	22.6 (-1.8)	24.6 (-4.5)	23.8 (-2.7)	22.6 (-0.3)	28.9 (-3.5)	19.6 (-1.3)	16.8 (-1.5)	25.6 (-3.1)
CogVideoX (5B)	18.9 (-0.8)	19.6 (-3.3)	21.3 (-3.2)	18.1 (-1.4)	25.6 (-1.3)	17.9 (-0.5)	12.6 (+0.8)	21.0 (-1.5)
Wan2.1 (1.3B)	30.1 (-3.5)	29.0 (-5.0)	28.5 (-2.8)	28.9 (-1.9)	35.2 (-3.6)	27.4 (-3.1)	18.7 (+0.0)	30.6 (-3.1)
Wan2.1 (14B)	33.6 (-3.5)	31.6 (-1.4)	<u>32.6</u> (-1.3)	32.9 (-3.5)	39.0 (-3.9)	29.6 (-1.3)	**23.7** (-1.5)	34.1 (-1.9)
MAGI-1	27.3 (-5.9)	26.6 (-4.0)	29.4 (-1.7)	28.5 (-5.1)	37.4 (-5.0)	30.4 (-5.0)	19.1 (-8.0)	32.6 (-5.3)
HunyuanVideo	26.7 (-0.5)	26.2 (+0.2)	28.1 (-0.1)	32.3 (-4.8)	36.1 (-1.9)	23.9 (+2.1)	<u>21.8</u> (-5.3)	30.0 (-0.9)
Cosmos (7B)	32.5 (-8.2)	32.8 (-9.3)	**33.6** (-8.9)	<u>36.1</u> (-10.6)	<u>40.1</u> (-10.6)	<u>31.6</u> (-10.0)	21.4 (-8.4)	<u>35.0</u> (-8.3)
Cosmos (14B)	**35.6** (-5.9)	<u>33.3</u> (-6.1)	31.7 (-3.9)	**37.4** (-5.9)	**40.9** (-9.6)	**32.5** (-5.3)	21.0 (-0.4)	**36.2** (-7.1)
Sora	28.6 (+2.4)	30.0 (-1.1)	29.9 (+0.2)	33.1 (+0.3)	37.0 (-1.1)	24.8 (+1.7)	16.7 (+2.6)	31.4 (+0.3)
Veo-2	<u>34.9</u> (-0.7)	**33.8** (-2.3)	31.7 (-2.7)	36.0 (-3.1)	38.4 (-1.5)	29.7 (+1.0)	19.5 (+2.5)	34.8 (-0.7)

textual evidence available to the judge.

Stage 3: LLM as a Judge. We use Gemini-2.5-Flash as the LLM judge. For each question, the judge receives all eight captions and the yes/no question; it does not see the original video or prompt – the questions alone carry the prompt's physical expectations. Because the eight captions emphasize different physical aspects, we count a question correct when *any* caption provides sufficient evidence for a "yes", so that a correct video is not penalized when a single caption happens to omit the relevant cue. Final performance is the proportion of questions answered correctly.

By decomposing evaluation into question generation, caption grounding, and LLM judging, our benchmark avoids conflating perception, grounding, and reasoning in a single VLM decision. The resulting scores measure whether generated videos contain captionable evidence for the physical behavior specified by language, making the benchmark scalable and its judgments interpretable.

For a complete end-to-end example, see App. F.

4 Experimental Setup

Models. We evaluate recent open-source and proprietary text-to-video (T2V) models, VideoCrafter2 (Chen et al., 2024), CogVideoX-2B/5B (Yang et al., 2025), Wan2.1-1.3B/14B (Team Wan et al., 2025), Cosmos-7B/14B (NVIDIA et al., 2025a), HunyuanVideo (Kong et al., 2025), LTX-Video (HaCohen et al., 2024), MAGI-1 (Sand.ai et al., 2025), Sora (Brooks et al., 2024), and Veo-2 (van den Oord and Roman, 2024). Open-source inference runs on a single NVIDIA A40 GPU; proprietary outputs are generated through their public interfaces.

Model-based difficulty grouping. We partition PhysVidBench into *medium*, *hard*, and *very hard* subsets to analyze how performance varies with physical-reasoning difficulty. We omit an "easy" tier since all retained prompts require nontrivial physical commonsense. As a difficulty proxy, we use the mean caption-grounded QA accuracy of four high-performing open-source models (Cosmos, HunyuanVideo, MAGI-1, Wan2.1), so that the estimate reflects difficulty shared across systems rather than any single model's behavior. Prompts scoring above 40%, 20–40%, and below 20% are assigned to the medium, hard, and very hard subsets (123, 160, and 100 prompts). We use these performance-derived groups as an analysis tool for comparing model behavior across increasingly challenging prompts.

5 Results and Discussion

Table 3 reports accuracy across the seven dimensions, and Figure 5 shows the distribution of per-prompt accuracies. The strongest open-source models, notably Cosmos-14B and Wan2.1-14B, obtain the highest averages, but performance is highly variable across prompts and dimensions. Propri-

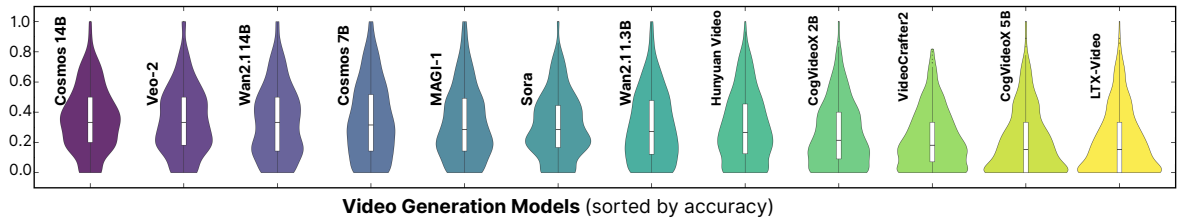


Figure 5: **Performance comparison of video generation models on PhysVidBench.** Each violin plot represents a distinct model and depicts the distribution of accuracy scores across all prompts.

Table 4: **Accuracy across model-based difficulty groups.** Values are caption-grounded QA accuracy percentages on upsampled prompts. The values inside the parentheses show the change when switching to base prompts.

Model	Medium	Hard	Very Hard
LTX-Video	41.2 (-1.6)	22.8 (-4.4)	9.1 (-1.2)
VideoCrafter2	44.5 (-4.3)	23.1 (-1.1)	10.7 (-1.6)
CogVideoX (2B)	49.8 (-7.8)	28.2 (-3.2)	11.6 (-0.6)
CogVideoX (5B)	39.2 (-2.7)	22.0 (-0.7)	11.3 (-1.1)
Wan2.1 (1.3B)	57.4 (-3.8)	33.8 (-4.7)	14.6 (-1.3)
Wan2.1 (14B)	61.5 (-4.8)	37.0 (-2.7)	18.0 (+0.2)
MAGI-1	60.5 (-10.7)	35.2 (-5.5)	16.8 (-3.2)
HunyuanVideo	52.6 (-4.5)	33.5 (-5.1)	15.9 (-4.8)
Cosmos (7B)	60.5 (-10.7)	38.8 (-10.3)	18.6 (-5.0)
Cosmos (14B)	59.2 (-13.1)	39.2 (-8.7)	22.2 (-2.2)

etary models such as Sora and Veo-2 are competitive in some settings but do not consistently dominate the benchmark. No evaluated model reliably handles the full range of everyday physical commonsense in PhysVidBench. Accuracy also differs across the model-based difficulty groups (Table 4, App. E).

Performance by physical dimension. Two trends recur. *Spatial Reasoning* and *Temporal Dynamics* are consistently among the hardest dimensions, even for Wan2.1-14B, Cosmos-14B, Sora, and Veo-2; these require preserving geometric relations, occlusion, event order, and temporal causality, which remain difficult despite high visual fidelity. *Object Properties & Affordances* yields the highest scores, which we read as evidence that current models depict recognizable objects and simple affordance cues more readily than they model relational, temporal, or causal structure. Visual plausibility alone is thus not sufficient for physical commonsense. Prompt-level bootstrap confidence intervals for the dimension-level results are reported in App. I. See App. G for qualitative results.

Effect of prompt enrichment. We compare base and enriched prompts to test whether making physical constraints explicit improves generation. Prompt enrichment yields modest, model-dependent mean gains, with paired bootstrap 95% confidence intervals excluding zero for six of the ten open-source models tested (Table A.8). The effect is not uniform. Sora declines under enrichment, possibly because its pipeline already performs internal prompt optimization, as discussed in VBench-2.0 (Zheng et al., 2025). Enrichment therefore provides model-dependent benefits but does not close the underlying physical-reasoning gap.

Effect of model scale. Larger models often score higher, but scale alone does not solve physical commonsense, and the trend is not monotonic. Within families, Wan2.1-14B improves over Wan2.1-1.3B, yet CogVideoX-2B matches or exceeds CogVideoX-5B on most dimensions, and smaller systems such as MAGI-1 and Wan2.1-1.3B can outperform larger ones such as VideoCrafter2 and CogVideoX-5B. We take this as a signal that factors beyond parameter count, such as architecture, training data, and supervision, matter for physical reasoning, though our setup does not isolate their individual contributions.

Human alignment of the evaluation pipeline. To validate the evaluator, we run a user study in which 15 participants watch 120 generated videos and answer the same yes/no questions used by the automatic judge, presented in randomized order. Participants see the video while the judge sees only captions, so the study measures how well caption-grounded reasoning recovers full-information human judgments. We report Pearson correlation with human responses separately for four generation models (Table 5). Across all four, our pipeline correlates most strongly with humans (0.45–0.69), outperforming every alternative. The best VLM judge, Qwen2.5-VL-72B-Instruct-AWQ (Bai et al., 2025b), trails throughout (0.21–0.47), and two video VLMs collapse entirely, answering "yes" to

Table 5: **Comparison of automatic evaluation methods.** Pearson correlation with human judgments on 120 videos. N/A denotes VLM collapse, where a model answers "yes" to nearly all questions. Our pipeline aligns with humans more strongly than prompted or specially fine-tuned VLM evaluators.

Evaluator	MAGI-1	Cosmos-14B	HunyuanVideo	Wan2.1-14B
Qwen2.5-VL-72B-Instruct-AWQ (Bai et al., 2025b)	0.47	0.36	0.42	0.21
Video-LLaVA (Lin et al., 2024)	N/A	N/A	N/A	N/A
InternVL3 (Zhu et al., 2025)	N/A	N/A	N/A	N/A
VideoScore (He et al., 2024)	-0.26	0.24	-0.09	-0.08
VideoPhy-2 AutoEval (Bansal et al., 2026)	0.23	-0.18	0.16	-0.06
Ours	**0.69**	**0.60**	**0.61**	**0.45**

nearly all questions. Notably, VLMs fine-tuned specifically for video evaluation (VideoScore (He et al., 2024), VideoPhy-2 AutoEval) do no better. Krippendorff's α ranges from 0.71 to 0.77 for items with at least three responses and from 0.75 to 0.85 for items with at least four responses. Our pipeline that separates evidence extraction from reasoning produces a more reliable evaluator of physical commonsense than prompted and fine-tuned VLMs. As complementary robustness checks, a manual audit of the same 120 human-study videos yields 98.5% event-level caption recall, while mismatching question sets and videos for Wan2.1-14B reduces the positive-response rate from approximately 34% to 4%. For more detailed analyses, see App. H and I.

Error-guided prompt refinement. Finally, we ask whether PhysVidBench can provide actionable feedback, not only aggregate scores. Following iterative self-refinement (Xue et al., 2025), we treat failed QA items as constraints for revising prompts without updating model parameters. Each round generates a video, the pipeline identifies violated physical requirements, and the next prompt adds targeted constraints (object attributes, interaction verbs, temporal cues, or explicit negatives). Unlike the fixed enriched prompts evaluated above, this procedure uses evaluation feedback to iteratively revise the prompt after generation. On 100 prompts, CogVideoX–2B and CogVideoX–5B improved from 21.6→32.7 (+11.1 percentage points) and 17.8→29.7 (+11.9 percentage points), respectively. This shows that our PhysVidBench evaluator can expose interpretable physical failures and guide prompt-level refinement.

6 Conclusion

We introduced PhysVidBench, a benchmark and language-grounded evaluation framework for physical commonsense in text-to-video generation, with an emphasis on tool use, object affordances, and causal interactions largely overlooked by prior benchmarks. By decomposing evaluation into captioning and caption-grounded question answering, it separates perception from reasoning and consistently aligns more strongly with human judgment than the prompted or fine-tuned VLM evaluators across all four evaluated generators. Across open and proprietary models, none reliably handles everyday physics, prompt enrichment helps only modestly, and scale does not close the gap. Our evaluator can also guide training-free prompt refinement.

Limitations

By design, our evaluator reasons over captions rather than raw video. This enables the perception–reasoning separation and traceability central to our approach, but makes evaluation dependent on caption quality: omitted information cannot be recovered by the judge. We mitigate this dependence through multi-perspective captioning (App. D) and robustness analyses. Replacing AuroraCap with Qwen2.5-VL-72B-Instruct-AWQ (Bai et al., 2025b) or Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a) yields highly consistent scores (App. I.2), while replacing the Gemini-Flash judge with DeepSeek nearly preserves model rankings (Spearman ρ=0.98; Sec. I.1). The protocol is also not intended for correctness that depends on continuous visual dynamics, such as wave profiles, flame motion, or turbulent mixing. Finally, PhysVidBench is smaller than some concurrent benchmarks, reflecting our emphasis on human-validated prompts and tool-mediated interactions. PIQA predominantly reflects Western household scenarios, leaving cross-cultural generalization as an open question.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Zechen Bai, Hai Ci, and Mike Zheng Shou. 2025c. Impossible videos. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 2458–2483. PMLR.

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

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

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

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. 2024. Video generation models as world simulators. OpenAI Technical Report.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. 2025. AuroraCap: Efficient, performant video detailed captioning and a new benchmark. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. 2025. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, volume 2025, pages 97959–98108.

Anastasis Germanidis. 2024. Runway Gen-3 Alpha: AI video generation with temporal control. https://runway.com/research/introducing-gen-3-alpha. Accessed May 10, 2025.

Google DeepMind. 2025a. Gemini 2.5 Flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash.

Google DeepMind. 2025b. Gemini 2.5 Pro. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro.

Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. 2025. ETVA: Evaluation of text-to-video alignment via fine-grained question generation and answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21299–21309.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. 2024. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*.

Haoran He, Yang Zhang, Liang Lin, Zhongwen Xu, and Ling Pan. 2026. Pre-trained video generative models as world simulators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4645–4653.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. 2024. 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*, pages 2105–2123. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023a. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.

Yuqi Hu, Longguang Wang, Xian Liu, Ling-Hao Chen, Yuwei Guo, Yukai Shi, Ce Liu, Anyi Rao, Zeyu Wang, and Hui Xiong. 2026. Simulating the real world: A unified survey of multimodal generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(9):11007–11026.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023b. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20406–20417.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024a. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631, Bangkok, Thailand. Association for Computational Linguistics.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024b. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, and 33 others. 2025. HunyuanVideo: A systematic framework for large video generative models. *Preprint*, arXiv:2412.03603.

Kuaishou Team. 2024. Kling. `https://klingai.kuaishou.com/`. Accessed May 12, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024a. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. 2024b. Sora: A review on background, technology, limitations, and opportunities of large vision models. *Preprint*, arXiv:2402.17177.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2023. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. In *Advances in Neural Information Processing Systems*, volume 36.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. 2025. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 43781–43806. PMLR.

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

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, and 59 others. 2025a. Cosmos world foundation model platform for Physical AI. *Preprint*, arXiv:2501.03575.

NVIDIA, Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Liang Feng, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth Gururani, Imad El Hanafi, Zekun Hao, Jacob Huffman, and 34 others. 2025b. Cosmos-Reason1: From physical common sense to embodied reasoning. *Preprint*, arXiv:2503.15558.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2234–2242, Red Hook, NY, USA. Curran Associates Inc.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang,

Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, and 20 others. 2025. MAGI-1: Autoregressive video generation at scale. *Preprint*, arXiv:2505.13211.

Jaskirat Singh and Liang Zheng. 2023. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and advanced large-scale video generative models. *Preprint*, arXiv:2503.20314.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. In *ICLR 2019 Workshop on Deep Generative Models for Highly Structured Data*.

Aäron van den Oord and Elias Roman. 2024. State-of-the-art video and image generation with Veo 2 and Imagen 3. https://blog.google/innovation-and-ai/models-and-research/google-labs/video-image-generation-update-december-2024/. Accessed May 11, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. 2024. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*.

Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. 2024. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025. 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)*, pages 18826–18836.

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. 2025. CogVideoX: Text-to-video diffusion models with an expert transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, Miami, Florida, USA. Association for Computational Linguistics.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. 2025. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *Preprint*, arXiv:2503.21755.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Appendix

In this supplementary material, we provide:

- A comparison of PhysVidBench with prior physical commonsense video benchmarks, namely PhyGenBench, VBench 2.0, VideoPhy-2, and Impossible Videos, with representative prompts and diversity statistics (App. A).
- A worked example showing that a strong off-the-shelf video metric fails to separate correct from incorrect generations of the same prompt (App. B).
- Our benchmark design, including prompt generation and upsampling templates, and the expanded seven-dimension physical commonsense ontology, with per-dimension definitions and example base vs. upsampled prompts (App. C).
- Our dense captioning and QA workflow, including dimension-targeted AuroraCap prompts, sample captions, and the full judge prompt (App. D).
- A taxonomy of physically grounded failure modes along three axes, Action Complexity, Scientific Principle, and Object Functionality, explaining *why* certain prompts are hard (App. E).
- A full end-to-end example evaluation on PhysVidBench (App. F).
- Qualitative examples contrasting successful and failed generations across models and dimensions, with key frames and QA verdicts, and the full specifications of the tested models (App. G).
- Details of our human evaluation study, including interface screenshots and human–LLM agreement (App. H).
- Robustness analysis of the evaluation pipeline, including cross-judge stability, captioner sensitivity, caption recall, a mismatched-video control, and prompt-level confidence intervals for dimension-level results, error-guided prompt refinement, and statistical significance analysis (App. I).

A Expanded Comparison with Existing Benchmarks

A.1 Representative Prompts from Each Benchmark

While prior video-generation benchmarks probe particular aspects of physical reasoning, none offer the breadth and task-oriented focus of PhysVidBench.

- **PhyGenBench** (Meng et al., 2025) targets isolated physical laws like buoyancy, stress, elasticity through scenarios such as a sponge being squeezed or a ball bouncing. Its prompts highlight single phenomena in passive settings and omit any agent-driven tool use or multi-step manipulations.
- **VBench 2.0** (Zheng et al., 2025) introduces a wide variety of scene types and evaluates models on spatial consistency, temporal coherence, and visual realism. However, it devotes limited attention to object affordances or the mechanics of purposeful interactions, testing where objects "are" rather than what agents can "do" with them.
- **VideoPhy-2** (Bansal et al., 2026) emphasizes motion plausibility and physical violations, e.g. a Segway traversing speed bumps or a gymnast back-flipping near a wall. Yet, these scenarios rarely involve purposeful tool handling or causal sequences of actions.
- **Impossible Videos** (Bai et al., 2025c) leverages visually striking, intentionally implausible phenomena such as endless pouring of liquid, and self-moving books to expose hallucinations in generation models. While effective at uncovering obvious physics failures, its surreal setups lack grounding in real-world, goal-driven tasks.

In contrast, **PhysVidBench** is built around grounded, goal-oriented prompts derived from the PIQA dataset. Each scenario requires agents to use or repurpose everyday objects in deliberate sequences that simultaneously engage multiple commonsense dimensions (force application, spatial fit, material properties, containment, and procedural causality). This affordance-centric design makes PhysVidBench uniquely suited for assessing whether text-to-video models truly "understand" how tools and materials behave in practical, manipulable tasks.

In the following, we list representative prompts from each benchmark to illustrate these contrasts.

PhyGenBench

Example Prompt 1: A cup of water is slowly poured out in the space station, releasing the liquid into the surrounding area.

Tests gravity absence in a synthetic environment; lacks real-world, goal-directed action or interaction with tools.

Example Prompt 2: A timelapse of a water-filled sponge being forcefully squeezed by hand,

with the pressure intensifying rapidly over time.

Shows material deformation under stress, but the action is passive. It does not involve affordance-based manipulation or multi-step use of the sponge.

Example Prompt 3: A vibrant, elastic rubber ball is thrown forcefully towards the ground, capturing its dynamic interaction with the surface upon impact.

Focuses on elasticity and rebound, but lacks physical goal structure, repurposed tool logic, or causal task sequences.

VBench 2.0

Example Prompt 1: A brown dog is on the left of an apple, then the dog moves to the right of the apple.

Tests spatial consistency from different perspectives but lacks physical interaction, causality, or manipulation involving materials or tools.

Example Prompt 2: A jar of peanut butter is opened in the space station, with the viscous liquid slowly dispersing.

Depicts fluid behavior in zero gravity but is passive and context-specific. There is no goal-directed manipulation or tool-based interaction.

Example Prompt 3: A person is drinking a glass of water, then they suddenly start cleaning the windows.

Shows a behavioral transition but omits the physical process involved in cleaning. The objects used and the steps required are unspecified and ungrounded.

VideoPhy-2

Example Prompt 1: A Segway bumps over a series of small speed bumps, the rider maintaining balance with slight adjustments.

Evaluates motion plausibility and equilibrium but lacks goal-directed object use, sequential interaction steps, or any reasoning about tool affordances.

Example Prompt 2: A backflip is performed close to a wall. The person carefully ensures their momentum keeps them clear from hitting the wall.

Involves physical trajectory reasoning but does not require interaction with any objects or manipulation of the environment for task completion.

Example Prompt 3: A backhoe digs a large hole in the ground, throwing up a mound of dirt.

Depicts mechanical tool usage, yet the prompt remains high-level and underspecified. It does not describe the procedural steps or physical constraints involved in the interaction.

Impossible Videos

Example Prompt 1: A person continuously pours water from a glass pot, yet remarkably the water level inside remains constant. This photo-realistic illusion defies natural physics as the endless stream flows without depleting the pot's contents. The scene takes place against a plain background, creating a focused view of this seemingly impossible phenomenon.

Intentionally violates conservation of mass for visual paradox, but lacks grounded, goal-driven interaction or affordance-sensitive manipulation.

Example Prompt 2: A metallic iron ball bounces off a polished marble floor in a physically unusual manner. Defying normal physics, each consecutive bounce reaches a greater height than the last, creating an eerily unnatural progression of increasing rebounds. The glossy surface of the floor reflects the ball's motion in this photo-realistic scene, set against what appears to be an indoor space.

Targets energy conservation violation in isolation. However, it contains no agent, object use, or causal physical interaction sequences grounded in real-world affordances.

Example Prompt 3: A book mysteriously opens by itself on a plain wooden desk surface, its pages flipping autonomously without any visible human

Table A.1: **Lexical diversity comparison across benchmarks.** PhysVidBench (base and upsampled variants) exhibits high diversity in verb lemmas, nouns, and action–object pairs relative to prior physics-oriented video benchmarks. These statistics provide a lexical proxy for the variety of physical interactions represented in the prompts.

Benchmark	# Prompt	Unique Verbs	Verbs per Prompt	Unique Nouns	Nouns per Prompt	Action–Object Pairs per Prompt	Lexical Density
PhyGenBench	160	78	0.488	239	1.494	0.79	0.55
PhyGenBench (explicit prompt)	160	232	1.45	408	2.55	2.91	0.57
VideoPhy-2 (train set)	3343	773	0.231	2111	0.631	1.43	0.59
VBench 2.0	1013	593	0.585	1282	1.266	1.75	0.55
VBench 2.0 (augmented)	1013	1195	1.18	2979	2.941	6.37	0.61
PhysVidBench (base)	383	262	0.684	708	1.849	2.03	0.60
PhysVidBench (upsampled)	383	625	1.632	1233	3.219	6.68	0.64

intervention or external force. The photo-realistic footage captures this unexplainable movement in a well-lit indoor setting, defying natural physics as the book cover and pages lift and turn on their own.

Demonstrates spontaneous motion but offers no agent–object interaction, mechanical force reasoning, or procedural physicality that would be necessary in task-driven video generation.

PhysVidBench (Ours)

Example Prompt 1: A slice of bread is pressed onto small glass fragments scattered on a surface, picking them up.

Involves soft material usage for delicate cleanup. Engages reasoning about texture, surface adhesion, and containment, grounded in everyday physical affordances.

Example Prompt 2: Duct tape is wrapped around a stuck light bulb in a socket, and a hand uses the tape to twist the bulb counter-clockwise, loosening it.

Demonstrates tool repurposing and torque application. Highlights grip mechanics, flexibility, and rotational force in a functional task context.

Example Prompt 3: A drink can is placed upright into the opening of a shoe resting on the floor of a car's passenger side.

Tests reasoning about shape compatibility, friction, and containment. Showcases improvised affordance in a constrained physical setup.

Complementing the lexical diversity statistics in Table A.1, PhysVidBench contains 62 of PhyGenBench's 78 unique verb lemmas (79.5%) and covers 75.5% of its verb occurrences. These lexical statistics provide an additional proxy for the variety of physical interactions represented in the prompts.

Cross-benchmark dimension coverage. Within PhysVidBench, the seven physical dimensions are assigned based on the manually reviewed QA items, with each prompt contributing to every dimension represented by at least one of its questions. For cross-benchmark comparison, we instead apply a common GPT-4o-mini-based multi-label classifier uniformly to the prompts of all benchmarks.

Under this common classifier, 97.7% of PhysVidBench prompts cover Object Properties & Affordances, compared with 27.5–59.1% in the comparison benchmarks. Coverage is 89.3% for Force and Motion, 71.0% for Material Interaction & Transformation, and 39.7% for Action & Procedural Understanding. Prompts jointly involving Object Properties & Affordances, Force and Motion, and Material Interaction & Transformation account for



(a) Physically correct (SA:3, PC:3)



(b) Physically wrong (SA:3, PC:3)



(c) Physically wrong (SA:3, PC:3)

Figure A.1: **Limitations of direct VLM-based evaluation for physical commonsense assessment.** While Semantic Adherence (SA) measures whether the generated video matches the input text prompt semantically, Physical Commonsense (PC) evaluates whether the generated content follows intuitive real-world physical laws. Existing direct scoring approaches often fail to distinguish semantically correct but physically implausible generations, assigning similarly high scores even when severe physical violations occur.

14.4% of PhysVidBench, compared with at most 2.0% elsewhere.

A.2 Dataset Diversity and Scale

PhysVidBench achieves broad conceptual coverage despite its compact size. Each prompt simultaneously probes several aspects of everyday physical reasoning such as affordances, causality, material response, and procedural manipulation, yielding higher supervision density than benchmarks built around isolated physical laws. As Table A.1 shows, the base and upsampled prompts exhibit high lexical diversity in verbs, nouns, and action–object pairs; we use this only as a proxy for the variety of real-world interactions, alongside the separate classifier-based coverage analysis above. Although the benchmark contains 383 base prompts, its effective evaluation scale is large: expanded across models, captions, and physics-grounded questions, it produces more than 42,000 QA instances per split, offering fine-grained diagnostic resolution in practice. This scale is consistent with established physics-oriented diagnostic datasets (PhyGenBench: 160 prompts; VideoPhy: 688 prompts), which likewise prioritize conceptual precision and interpretability over raw quantity. PhysVidBench follows the same design philosophy, remaining computationally feasible for the community while still rich enough to meaningfully evaluate physical commonsense in T2V models.

B Limitations of Direct VLM Scoring for Physical Commonsense

Direct VLM-based scoring is appealing because it is simple: a video and a prompt are passed to a vision-language model, which returns scalar scores for semantic adherence and physical correctness. Physical commonsense, however, is not merely a visual recognition problem. It requires verifying whether the video actually supports the causal and procedural consequences the prompt implies. Direct scoring is therefore vulnerable to shortcut judgments grounded in surface-level visual plausibility, object presence, or general prompt alignment.

Figure A.1 illustrates this failure mode in a controlled setting. We compare generations that share the same semantic context: a person uses a stone to hit a nail on a wooden surface. In the physically plausible case, the stone contacts the nail and the nail is gradually driven into the wood. In the physically implausible cases, the semantic scene is still recognizable, but the physical interaction breaks down: the stone either passes through the nail in an impossible manner or melts before producing the expected mechanical effect. Despite these severe physical violations, the direct VLM scorer (VideoPhy 2 Autoeval) assigns the same high scores to all cases, including both the plausible and impossible generations (SA:3, PC:3). This suggests that the scorer recognizes the objects, action context, and

overall visual scene, but fails to distinguish whether the required physical interaction actually occurs.

This behavior is especially problematic when direct VLM scores are used as training targets or reward signals. If a model is trained only to match scalar human or VLM scores, it may learn shortcuts that correlate with physical correctness without truly evaluating physical causality. For example, stronger video generation models often produce sharper, more coherent, and more realistic-looking videos; these same models are also less likely to contain obvious physical artifacts. As a result, visual quality can become spuriously correlated with physical quality. A direct evaluator trained on such scores may therefore learn to reward high visual fidelity, temporal smoothness, or object recognizability, rather than the actual satisfaction of the physical constraints implied by the prompt. In this case, the evaluator may appear correlated with human physical-quality judgments at the dataset level, while still failing on counterfactual examples where semantic alignment and visual plausibility are held fixed but physical correctness differs.

Our method is designed to reduce this shortcut risk. Instead of asking a VLM to directly assign a single physical correctness score, PhysVidBench decomposes evaluation into captioning and caption-grounded question answering. The video is first converted into explicit textual evidence, and then an LLM answers targeted yes/no questions about the expected physical consequences. In the example in Figure A.1, the evaluator does not simply ask whether the scene shows a person, a stone, and a nail. It checks whether the stone remains solid, whether it contacts the nail, whether the nail moves into the wood after impact, and whether the observed sequence supports the intended causal effect. If the caption evidence does not entail these physical events, the corresponding questions are answered negatively.

This decomposition makes evaluation more interpretable and less reliant on a single opaque judgment. Direct scoring can collapse a physically correct video and a physically impossible one onto the same score; caption-grounded QA instead pinpoints which physical requirement is missing. PhysVidBench thus provides a more reliable signal for physical commonsense, especially when visually realistic generations still violate basic causal, material, or interaction-level constraints.

C Benchmark Design and Prompt Construction

C.1 Prompt Generation and Upsampling

Our benchmark construction pipeline begins with the PIQA dataset, utilizing its extensive repository of physically grounded scenarios. PIQA contains roughly 21K total examples. Because the test split lacks public ground-truth labels, our processing pipeline uses the labeled train and validation splits, yielding more than 17K correct goal-solution pairs before filtering, from which 383 final prompts are retained. Our pipeline then proceeds through three stages: (1) *Initial filtering*; (2) *Video prompt generation*; and (3) *Prompt upsampling*.

Stage 1: PIQA Dataset and Initial Filtering. We use Gemini 2.5 Pro model to select only those PIQA scenarios that involve secondary or repurposed tool use, those that inherently require physical-commonsense reasoning. This filtering yields a compact set of complex, interaction-centric tasks ideal for text-to-video evaluation. Below is an example prompt illustrating the instructions provided to Gemini 2.5 Pro in this stage.

Video Generation Pipeline – Stage 1: PIQA Dataset and Initial Filtering

I will provide a series of examples from the PIQA dataset where common household objects or tools are used either in their primary/intended way or in repurposed/alternative ways (i.e., "life hacks"). Your task is to determine whether a given PIQA goal–solution pair reflects the primary usage of an object or not.

Definition: If the object is being used in its intended, standard, or conventional way, respond with: No
If the object is being used in an alternative or repurposed way (i.e., not its typical function), respond with: Yes

You are given a list of goal–action pairs (e.g., "Water Bottle → Egg Separator", "Credit Card → Ice Scraper") which illustrate how PIQA questions often involve creative or secondary uses of common items. These examples represent secondary usage. This list can guide your understanding of what constitutes repurposing. You will now be given a PIQA goal and two solutions. Based

on your understanding of object affordances and typical usage. Determine only whether the goal/solution pair represents a secondary use of an object. Write only the answer, Yes or No. Do not explain or describe the reasoning.

Here are examples from the PIQA dataset where common objects or tools are used for a purpose other than their primary, intended function (i.e., alternative uses or "life hacks"):

Water Bottle → Egg Separator:
Goal: To separate egg whites from the yolk using a water bottle...
Action: Squeeze the water bottle and press it against the yolk. Release, which creates suction and lifts the yolk.

Hair Net → Vacuum Filter:
Goal: How do I find something I lost on the carpet?
Action: Put a hair net on the end of your vacuum and turn it on. (To catch small items before they go into the vacuum bag).

These examples demonstrate the kind of creative repurposing and understanding of object affordances beyond their primary function that the PIQA dataset aims to capture.
Write only the answer for each goal with comma separator: Yes or No. Do not explain or describe the reasoning.

Give the answer for this Goal Solution pair:
Goal: How can I paint stars on a black background?
Solution: Use an old toothbrush covered in white paint and flick at the background.

Stage 2: Video Prompt Generation. We use the Gemini 2.5 Pro model to transform each filtered goal–solution pair into a concise, visually explicit video generation prompt. This process produces 383 distinct *base* prompts, each outlining a self-contained scenario with clear, physically grounded actions. Below is an example prompt illustrating the instructions provided to Gemini 2.5 Pro in this stage.

Video Generation Pipeline – Stage 2: Video Prompt Generation

You are given question–answer pairs from the PIQA dataset. Each pair consists of a physical goal (what someone wants to achieve) and a solution (how to achieve it). Your task is to write a **realistic and physically plausible video generation prompt** that visually demonstrates the provided solution.

Important Instructions:
- Only generate a short, visualizable **video prompt** that clearly shows the solution being carried out.
- The prompt should describe **a realistic short video clip** (under 10 seconds).
- You do **not** have to begin each prompt with "A person". Be natural — use whatever fits best (e.g., "A dog...", "Hands...", "Someone...", "Two people...", "The object...").
- If the solution is not visualizable or does not reflect meaningful physical interaction or commonsense action, **skip it** (do not generate a prompt).
- Be **specific** and **physically grounded** — actions should be directly observable (e.g., cutting, pouring, lifting, placing).
- Do not repeat the goal or the original solution text.

Example Conversions:

Goal: You want to dry your wet hands.
Solution: Use a towel
Video Prompt for Solution: Wet hands are rubbed against a cotton towel.

Goal: You want to keep a door open.
Solution: Use a doorstop
Video Prompt for Solution: A door is held open by sliding a rubber doorstop underneath.

Goal: You want to get rid of wrinkles in a shirt.
Solution: Use an iron
Video Prompt for Solution: An iron glides over a wrinkled shirt on an ironing board, flattening it.

Now, for each of the following goal–solution pairs from the PIQA dataset, return a single line in the format below **only if it qualifies**:

PIQA Input:
Goal: How can I paint stars on a black background?
Solution: Use an old toothbrush covered in white paint and flick at the background.

Stage 3: Prompt Upsampling. We use the Gemini 2.5 Pro model to enrich each *base* prompt with fine-grained physical details such as explicit force application, material behaviors, and spatial constraints, while preserving the original scene scope. This controlled upsampling yields 383 *upsampled* prompts that surface deeper commonsense cues without introducing new objects or outcomes. Below is an example prompt illustrating the instructions provided to Gemini 2.5 Pro in this stage.

Video Generation Pipeline – Stage 3: Prompt Upsampling

You are a reasoning-augmented AI model tasked with upsampling video generation prompts using knowledge of physical commonsense and object affordances, while strictly preserving the original scene boundaries. Given a simple prompt, you must rewrite it into a richer, more physically grounded version by:

- Elaborating on object properties (e.g., material, size, rigidity, weight) that affect how actions are performed
- Clarifying affordances (e.g., what can be held, moved, poured, tied) based on those properties
- Expanding the action description in a physically realistic, causally plausible sequence
- Including any relevant environmental or contextual conditions (e.g., gravity, surface type)
- Do not introduce any new objects or tools not mentioned in the original prompt
- Do not describe outcomes or events that happen after the main action (no result, consequence, or future speculation)
- Your goal is to make the prompt maximally useful for video generation models that rely on physical reasoning, without hallucinating or extending beyond the original scope.

Example Input:
"A person places coffee filters inside the cup holders of a car."

Upsampled Output:
"A person carefully pushes white paper coffee filters, each lightweight and slightly crinkled, into the round plastic cup holders on the car's center console. The filters conform to the shape of the holders due to their flexible, foldable structure. The car interior is clean and the cup holders are empty, allowing the filters to settle flush against the bottom surface."

Only return the upsampled prompt as output. Do not explain your reasoning or include anything else in your answer.

Prompt: A person dips an old toothbrush in white paint and flicks it onto a black background, creating stars.

Manual validation was performed by four authors using a shared review form and written criteria covering visualizability, tool use, affordance relevance, and physical commonsense. Each candidate was independently reviewed by two authors, and disagreements were resolved through discussion before inclusion.

Altogether, PhysVidBench comprises 766 unique prompts (383 base + 383 upsampled), enabling a comprehensive evaluation of physical commonsense reasoning in text-to-video models. Representative examples illustrating the outcome of each stage of our pipeline are shown below.

Example 1

Stage 1: PIQA Dataset and Initial Filtering
Goal: How can I paint stars on a black background?
Solution: Use an old toothbrush covered in white paint and flick at the background.

Stage 2: Video Prompt Generation
A person dips an old toothbrush in white paint and flicks it onto a black background, creating stars.

Stage 3: Prompt Upsampling
A person firmly grips the handle of an old, worn toothbrush, its plastic slightly yellowed with age. They dip the frayed bristles into a can of thick, white paint. The paint clings to the bristles due to its viscosity. With a controlled wrist motion, they flick the paint-laden toothbrush towards a large, matte black canvas backdrop. Tiny droplets of white paint

detach from the bristles and spray outwards, propelled by the force of the flick. The droplets travel through the air, influenced by gravity, before adhering to the vertical surface of the canvas.

Example 2

Stage 1: PIQA Dataset and Initial Filtering
Goal: When I am driving an older car with no cupholders where can I place my beverage?
Solution: Use an old shoe, place the drink container in the place where your foot enters the shoe and it will help stabilize it.

Stage 2: Video Prompt Generation
A drink can is placed upright into the opening of a shoe resting on the floor of a car's passenger side.

Stage 3: Prompt Upsampling
A cylindrical aluminum drink can, rigid due to its metal construction and filled with liquid giving it noticeable weight, is carefully maneuvered vertically downwards. It is inserted, base-first, into the flexible fabric opening of a casual shoe which is resting flat on its firm sole upon the textured floor surface of the car's passenger side. The can maintains its upright orientation under gravity as its smooth, flat circular base makes stable contact with the interior insole surface deep within the shoe's cavity.

Example 3

Stage 1: PIQA Dataset and Initial Filtering
Goal: How do I find something I lost on the carpet?
Solution: Put a hair net on the end of your vacuum and turn it on.

Stage 2: Video Prompt Generation
A person attaches a hair net to the end of a vacuum cleaner and uses it to search for a small object on a carpet.

Stage 3: Prompt Upsampling
A person stretches the elastic rim of a fine, white nylon hair net, expanding its opening, and carefully fits it over the circular nozzle of a dark gray plastic vacuum cleaner hose. The net conforms to the hose's shape, held in place by its own tension. The vacuum cleaner is then used to gently hover over a low-pile carpet, its fibers slightly compressed by the vacuum's suction. The person guides the nozzle in slow, overlapping passes, the hair net acting as a barrier to prevent small objects from being sucked into the machine.

C.2 Physical Commonsense Dimensions and Sample Prompts

We present each of the seven physical commonsense dimensions defined in our benchmark, with representative base and upsampled prompts.The prompts shown under each category were chosen because they most strongly emphasize that form of reasoning, but PhysVidBench prompts are not confined to a single dimension. In practice, most scenarios engage several at once. A prompt may combine force application with spatial configuration, or material properties with temporal sequencing. This multi-dimensional grounding mirrors the complexity of real physical tasks and distinguishes PhysVidBench from benchmarks that test physical concepts in isolation.

Fundamental Physics: ***The foundational laws governing energy, causality, equilibrium, and changes of physical state***

Base Prompt: A flashlight lying on a dark floor illuminates tiny shards of broken glass by casting long shadows from them.
Upsampled Prompt: A cylindrical metal flashlight rests horizontally on a dark, flat floor, its switch in the 'on' position. A focused beam of bright light emits from its lens, traveling parallel and close to the floor surface. This low-angle light strikes multiple tiny, sharp-edged, transparent shards of broken glass scattered across the floor. Because the light source is nearly level with the floor, even the small vertical profile of each glass shard is sufficient to block the light path, causing disproportionately long, thin shadows to stretch out behind each shard, starkly visible against the surrounding darkness of the floor.

Base Prompt: A hand pours a small amount of white flour into a larger bowl of water and begins stirring with a spoon.
Upsampled Prompt: A human hand directs a small stream of fine, white, lightweight flour powder downwards due to gravity, letting it fall onto the surface of still water held within a larger, rigid bowl resting on a flat surface. The flour particles make contact with the liquid. The hand then

grips the rigid handle of a spoon, submerges the spoon's end into the water where the flour landed, and initiates a stirring motion, moving the spoon through the water and the suspended flour particles.

Base Prompt: A plastic bag containing meat sits on an upside-down aluminum pot, with another pot filled with water placed directly on top of the bagged meat.

Upsampled Prompt: A flexible, translucent plastic bag, visibly holding a weighty piece of raw meat which causes the thin material to bulge slightly, is positioned on the flat, circular, upturned bottom of a lightweight yet rigid aluminum cooking pot resting stably upside-down on its rim. Placed directly centered on top of this bagged meat rests another rigid cooking pot, its flat bottom making full contact; this second pot contains a significant volume of water, adding substantial downward weight due to gravity. The pliable plastic bag is compressed between the base of the top, water-filled pot and the upturned base of the bottom pot, conforming slightly to the shapes above and below it while transferring the load.

Object Properties & Affordances: ***The study of inherent object attributes, including material composition, rigidity, softness, and functional affordances like containment or support***

Base Prompt: A stainless steel bucket is placed upside down over a small ant crawling on the ground.

Upsampled Prompt: A rigid, opaque stainless steel bucket, positioned upside down with its open circular rim facing downwards, is carefully lowered vertically towards the ground. The bucket descends directly over a small, dark ant actively crawling across the relatively flat earth surface. The bucket's descent continues until its metal rim makes firm contact with the ground, encompassing the area occupied by the crawling ant.

Base Prompt: A hand rubs a piece of bread back and forth over pencil writing on paper.

Upsampled Prompt: A human hand grips a piece of soft, somewhat pliable white bread, likely without crusts, between its fingers. Positioned over a sheet of standard writing paper lying flat on a hard surface, the hand methodically rubs the bread's yielding, slightly textured underside back and forth. This motion occurs directly over an area marked with fine, gray pencil writing, causing the bread's surface to make repeated contact with the graphite lines on the paper under consistent, light pressure.

Base Prompt: Hands fold one short end of a credit card up and the other short end down, then place a smartphone onto the upward fold.

Upsampled Prompt: A pair of hands holds a standard-sized, thin, rectangular plastic credit card. Applying pressure with fingers, the hands bend one short edge of the rigid-yet-flexible card sharply upwards along its width, creating a defined crease. Then, the hands bend the opposite short edge sharply downwards along its width, forming a second parallel crease, configuring the card into a small S-shaped structure. Next, the hands carefully lift a smooth, relatively heavy, rigid rectangular smartphone. Aligning the bottom edge of the phone with the small, upward-facing plastic ledge created by the first fold on the credit card, the hands gently lower the weight of the smartphone onto this narrow support.

Spatial Reasoning: ***The ability to interpret and infer spatial relations, including position, geometry, occlusion, orientation, and fit between objects in a scene***

Base Prompt: The tip of a spoon presses a square outline into a slice of bread, then the back of the spoon pushes down the center.

Upsampled Prompt: The hard, pointed metal tip of a rigid spoon presses sequentially into the soft, porous surface of a slice of bread, tracing a square outline by creating shallow indentations. The bread yields slightly under the focused pressure of the tip. Subsequently, the smooth, curved underside (back) of the spoon's bowl is centered within the traced square. Applying downward force, the broader surface of the spoon back pushes into the bread, compressing the central area significantly more deeply than the initial outline, creating a distinct depression within the square boundary on the yielding bread slice.

Base Prompt: Crushed egg shells and water are

shaken vigorously inside a clear bottle.
Upsampled Prompt: Inside a clear, rigid bottle containing clear liquid water, numerous small, lightweight, brittle fragments of crushed egg shells are forcefully agitated. The bottle undergoes vigorous, rapid shaking motions, causing the water to slosh violently and carry the sharp-edged shell fragments in turbulent, swirling patterns within the confined space. The mixture of water and shell fragments repeatedly impacts the inner surfaces of the bottle due to the applied force.

Base Prompt: A hand pushes a plastic drinking straw upwards through the bottom of a strawberry, popping the green stem out the top.
Upsampled Prompt: A human hand firmly grips the lower end of a thin, cylindrical, rigid plastic drinking straw. The straw's end is pressed against the pointed bottom tip of a ripe, red strawberry, which is soft enough to be pierced but firm enough to maintain its shape. The hand applies a continuous, steady upward force along the axis of the straw. Due to its rigidity and the applied pressure, the straw penetrates the yielding flesh of the strawberry, moving vertically upwards through its center. As the leading end of the straw travels inside the fruit towards the wider top where the green, leafy stem is attached, it pushes against the stem's base from below. This focused upward force exerted by the straw dislodges the relatively light stem structure, causing it to pop out from the top surface of the strawberry.

Temporal Dynamics: ***The domain addressing timing, sequencing, and the causal structure of events over time, such as ordering, delays, and waiting***

Base Prompt: A plastic bucket is placed upside down on someone's head like a hat.
Upsampled Prompt: A lightweight, rigid plastic bucket, oriented upside down so its solid bottom faces upwards, is carefully lowered onto a person's head. The open circular rim of the bucket makes contact and rests stably upon the crown of the head, balanced there by gravity. The bucket's inherent rigidity allows it to maintain its shape, covering the top of the head like a large, hollow hat.

Base Prompt: A child places their hand on a large circle sticker affixed to the side of a car in a parking lot.
Upsampled Prompt: A small child extends their arm and carefully places their soft, open hand flat onto the smooth surface of a large, circular sticker. The thin, flexible sticker is firmly adhered by its backing to the hard, vertical metal side panel of a stationary car. The car is situated on the level asphalt ground of an outdoor parking lot. The child's palm and fingers press lightly against the sticker's surface, conforming slightly to the car's panel curvature beneath it.

Base Prompt: A hand squeezes toothpaste onto a cloth and rubs it in circles over a hazy car headlight, leaving a thin white coating.
Upsampled Prompt: A human hand squeezes a dollop of viscous, opaque white toothpaste directly onto a section of a soft, flexible fabric cloth. The paste adheres to the cloth's surface. The hand then presses this part of the cloth firmly against the hard, smooth, but visibly hazy surface of a car's plastic headlight cover. Maintaining steady pressure, the hand rubs the cloth in continuous circular motions over the hazy area. The pliable cloth conforms to the headlight's curvature, spreading the toothpaste evenly as it's rubbed, leaving behind a thin, uniform white coating on the headlight surface.

Action & Procedural Understanding: ***The comprehension of structured, goal-driven sequences of actions involving procedural steps toward task completion***

Base Prompt: Hands press a piece of clay over the joined edges of pieces on top of a bowl.
Upsampled Prompt: Human hands hold a separate lump of soft, malleable, slightly damp clay. The hands carefully position this clay lump directly over the visible seam where the edges of several harder, pre-joined clay pieces meet. These pieces rest securely on the upward-facing surface of a rigid ceramic bowl. Fingers then apply firm, steady downward pressure onto the soft clay lump, causing it to flatten and spread smoothly across the underlying seam, adhering to the surfaces of the joined pieces due to its pliable and slightly sticky nature.

Base Prompt: A cloth damp with olive oil is wiped over a metal key. The key is then inserted smoothly into a keyhole.
Upsampled Prompt: A soft, absorbent cloth, damp and glistening slightly with viscous olive oil, is wiped firmly across the surfaces of a rigid, metallic key. The pliable cloth presses into the key's teeth and grooves, coating the hard metal with a thin, lubricating film of oil. Immediately after, the oil-coated metal key is precisely aligned with the opening of a metal keyhole and inserted with a steady, linear motion, sliding effortlessly into the lock mechanism due to the oil reducing friction between the contacting metal surfaces.

Base Prompt: Hands carefully pour a green liquid over the back of a spoon into a shot glass containing a brown liquid, forming a distinct layer. Then, a cream liquid is poured similarly on top.
Upsampled Prompt: Hands carefully position a rigid metal spoon, its convex back facing upwards, just above the surface of a dark brown, relatively dense liquid resting at the bottom of a small, clear glass shot glass. With precise control, a vibrant green liquid, less dense than the brown liquid, is steadily poured onto the spoon's curved back. The spoon's smooth surface guides the green liquid's flow, allowing it to gently cascade onto the brown liquid's surface, minimizing turbulence and forming a distinct, level layer without significant mixing. Following this, the hands adjust the spoon's position slightly upwards, holding it just over the newly formed green layer. Then, an opaque, cream-colored liquid, possessing an even lower density than the green liquid, is poured using the same technique over the back of the spoon, flowing down gently to create a third, clearly separated layer resting atop the green one within the confines of the shot glass's rigid walls.

Material Interaction & Transformation: *The study of how materials respond to external forces or processes, including transformations like melting, freezing, breaking, or chemical change*

Base Prompt: Hands hold a smartphone inside a clear ziploc bag, tapping the screen while raindrops fall on the bag.
Upsampled Prompt: Human hands firmly grip the sides of a clear, flexible plastic ziploc bag, holding it steady. Inside the sealed bag rests a rigid, rectangular smartphone with its glass screen visible. A finger repeatedly taps onto the thin, pliable plastic surface directly overlying the phone's touch-sensitive screen, the pressure transmitting through the material. Simultaneously, small water droplets, falling under gravity, land and bead up on the exterior surface of the waterproof bag.

Base Prompt: Water flows from a sink faucet into the scoop of a dustpan, travels down its hollow handle, and pours into a bucket sitting beside the sink.
Upsampled Prompt: Clear water flows downwards in a steady stream from the metal nozzle of a sink faucet, driven by gravity. The stream lands directly onto the rigid, concave surface of a plastic dustpan's scoop, which is positioned underneath the faucet. Due to the angle of the dustpan, the accumulating water is channeled towards the opening at the base of the scoop leading into its hollow handle. The water then travels downwards through the enclosed tubular space within the handle. The open end of the handle is positioned directly over the wide opening of a sturdy plastic bucket resting stationary on the floor beside the sink, allowing the water exiting the handle to fall vertically into the bucket.

Base Prompt: A golf ball is placed on a countertop; it rolls slightly, then the counter is subtly adjusted, and the ball comes to a stop.
Upsampled Prompt: A small, dense, hard golf ball with a characteristic dimpled texture is gently placed onto a large, flat, smooth countertop surface. Because of its perfectly spherical shape and the countertop potentially being minutely uneven, the heavy ball starts to roll slowly across the rigid surface. Subsequently, the countertop experiences a subtle, barely perceptible shift in its orientation, adjusting its levelness. This slight change counteracts the initial impetus for rolling, altering the gravitational pull along the surface, and the ball, influenced by friction and the new level plane, decelerates smoothly until it comes to a complete standstill upon the countertop.

Force and Motion: ***The domain focused on physical interactions governed by forces, including motion, pushing, pulling, lifting, and properties like inertia***

Base Prompt: A strong magnet is slowly moved across a drywall surface, suddenly snapping and sticking to the wall over a stud location.
Upsampled Prompt: A dense, rigid metallic magnet, possessing a strong magnetic field, is held close to a vertical, painted drywall surface. It is guided slowly and steadily in a continuous motion across the slightly textured face of the wall. As the path of the strong magnet crosses over the position of a hidden ferromagnetic metal stud located just behind the thin drywall layer, the attractive magnetic force increases sharply and suddenly. This abruptly intensified pull overcomes the controlled movement, causing the heavy magnet to accelerate rapidly towards the wall, making firm contact and sticking securely against the drywall surface directly above the unseen stud's location due to the powerful magnetic attraction.

Base Prompt: Soft tortillas are draped and wedged between the bumps of an upside-down muffin pan that has been sprayed with oil.
Upsampled Prompt: Soft, pliable, circular tortillas, thin and flexible due to their material properties, are draped over the rigid structure formed by an upside-down metal muffin pan. Gravity pulls the lightweight tortillas downwards. They are then carefully pressed and wedged into the concave spaces between the protruding, rounded bumps of the pan. The tortillas bend and conform to the curved shape of these spaces. The pan's surface, both bumps and crevices, is coated in a visible layer of slick cooking oil, facilitating the placement and slight sliding of the tortillas as they are settled into position.

Base Prompt: A hand places a colorful silicone cupcake liner into the empty cup holder of a car's center console.
Upsampled Prompt: A hand gently pinches a colorful, flexible silicone cupcake liner between thumb and forefinger. The liner, lightweight and pliable with distinct fluted sides, is carefully lowered vertically into an empty, cylindrical cup holder recessed into the car's rigid plastic center console. The liner slides smoothly against the inner wall and comes to rest flat against the bottom surface of the holder, its flexible structure allowing it to fit snugly within the defined space.

D Evaluation Details

To evaluate physical commonsense on PhysVidBench, we employ a three-stage pipeline: (1) generate targeted yes/no questions from each upsampled prompt, (2) produce multi-perspective dense captions for the corresponding video, and (3) use an LLM to answer those questions based solely on the captions.

Stage 1: Physics-Grounded Question Generation. For each upsampled prompt, we prompt Gemini 2.5 Pro to craft precise yes/no questions that probe our seven physical-commonsense dimensions (e.g., "Did the person apply sufficient force to separate the objects?"). Every question is guaranteed to have a "Yes" answer and is answerable by inspecting the prompt alone. We fix and store this question set for downstream evaluation of each model's video output. Below is an example prompt used in this stage.

Sample Prompt for Evaluation – Stage 1: Physics-Grounded Question Generation

Generate as many diverse, concrete yes/no questions as possible based only on the provided prompt. Each question must:

- Be specific and answerable solely from the prompt (no outside assumptions)
- Focus on physical or procedural understanding, not abstract or emotional reasoning
- Be non-redundant — do not repeat similar ideas or reword the same question
- Cover a wide range of physical reasoning categories
- Be phrased so that the physically correct answer is Yes
- If a question relates to multiple reasoning categories, include all applicable types

Physical Reasoning Categories: (Choose these types, write only the category like Fundamental Physics, Object Properties etc.)

- **Fundamental Physics:** energy, causality, equilibrium, state change
- **Object Properties & Affordances:** material

type, rigidity, softness, containment
- **Spatial Reasoning:** fit, position, occlusion, geometry, orientation
- **Temporal Dynamics:** ordering, timing, waiting, delays
- **Action & Procedural Understanding:** goal-directed behavior, methodical steps
- **Material Interaction & Transformation:** melting, freezing, breaking, chemical change
- **Force and Motion:** pushing, pulling, lifting, inertia

Output Format:
- Questions and Answers:
- Q: <yes/no question>?
- A: Yes
- Type: <Reasoning Category 1, Reasoning Category 2, ...>

You must return:
- A rich, diverse set of yes/no questions
- Each formulated so that its correct answer is Yes
- Each labeled with all relevant reasoning categories, separated by commas

Start with the following prompt:
Prompt: Person firmly grips the handle of an old, worn toothbrush, its plastic slightly yellowed with age. They dip the frayed bristles into a can of thick, white paint. The paint clings to the bristles due to its viscosity. With a controlled wrist motion, they flick the paint-laden toothbrush towards a large, matte black canvas backdrop. Tiny droplets of white paint detach from the bristles and spray outwards, propelled by the force of the flick. The droplets travel through the air, influenced by gravity, before adhering to the vertical surface of the canvas.

Stage 2: Dense Video Captioning. For each generated video, we use the AuroraCap model (Chai et al., 2025) to generate eight distinct captions: one general-purpose caption and seven dimension-specific captions (one per commonsense dimension). For instance, to highlight spatial relationships, we use the prompt "Describe the spatial layout and object positions…"; to surface force interactions we use "Describe how forces are applied and resisted . . . " Each prompt ends with an instruction to report only directly observed content, which is intended to reduce unsupported inference while eliciting complementary perspectives. Below is the full template we provide to AuroraCap in this stage.

Prompt for Evaluation – Stage 2: Dense Video Captioning

- Describe the video in detail.
- Describe the physical principles at work in the video, such as energy transfer, causal relationships, balance or imbalance, and any visible changes in physical state and etc. (e.g., solid to liquid).
- Describe the main objects in the video focusing on their material properties (e.g., rigid, soft, metallic), and what actions they allow or prevent and etc. (e.g., can be squeezed, can contain something).
- Describe the spatial layout and relationships in the video: object positions, orientations, fit between shapes, occlusions, and how geometry affects interactions and etc..
- Describe the sequence and timing of events in the video, including any delays, waiting periods, or causal orderings between actions or state changes and etc..
- Describe the actions performed in the video, focusing on the goals of the agent, the order of steps taken, and whether the actions appear intentional or methodical and etc..
- Describe how materials change or interact in the video: melting, freezing, breaking, mixing, or undergoing chemical or physical transformations and etc..
- Describe how forces are applied in the video (e.g., pushing, pulling, lifting), how objects respond (e.g., acceleration, resistance), and any indications of inertia or physical resistance and etc..

Only describe what can be directly observed in the video. Do not make assumptions or include external knowledge that is not visually confirmed.

Stage 3: LLM as a Judge. We assemble a single text prompt that includes: (a) all yes/no questions generated in Stage 1, and (b) the eight dense captions produced in Stage 2. We pass this prompt to Gemini-2.5-Flash-Preview-04-17 (Google DeepMind, 2025a), instructing it to respond only with "Yes" or "No" for each question based solely on the provided captions. Below, we present the exact Stage 3 prompt used to evaluate the sample video generated by the Wan2.1 (14B) model from the

Stage 1 prompt. This prompt combines the eight dense captions produced in Stage 2 with the corresponding physical-commonsense questions from Stage 1, illustrating how the LLM integrates multi-perspective textual evidence and targeted queries during evaluation.



Example Prompt for Evaluation – Stage 3: LLM as a Judge

You are given 8 captions describing different aspects of the same video. Answer "Yes" ONLY if at least one caption supports it, otherwise answer "No".

Caption 1: The video captures a moment where a person is using a toothbrush to apply a white substance, likely toothpaste, onto a dark surface. The toothbrush is held in a way that suggests the person is in the process of brushing, with the bristles making contact with the surface. The substance is being applied in a way that creates a textured effect, with some areas appearing more saturated than others. The background is a solid, dark color, providing a stark contrast to the bright white of the toothpaste. The lighting is even, allowing for clear visibility of the toothbrush and the substance being applied. The focus is on the action of brushing, with the background remaining unchanged, emphasizing the simplicity of the scene.

Caption 2: The video captures a moment where a toothbrush is being used to apply a white substance, likely toothpaste, onto a dark surface. The energy transfer is evident as the bristles of the toothbrush create a splatter effect, with the white toothpaste dispersing in a way that suggests a mix of solid and liquid components. The balance of the scene is slightly tilted, with the toothbrush being the focal point, while the dark surface provides a contrast that highlights the action. The physical state of the toothbrush and toothpaste is clearly observable; the toothbrush is in motion, and the toothpaste is in a semi-liquid state, transitioning from a solid to a liquid form as it spreads.

Caption 3: The main objects in the video are a toothbrush and a paintbrush. The toothbrush, with its soft bristles, is designed for cleaning teeth, allowing it to gently scrub surfaces. The paintbrush, with its metallic handle, is used for applying paint, preventing the paint from dripping and allowing for controlled application. The actions they allow are brushing and painting, respectively, while the toothbrush's soft bristles prevent abrasion and the paintbrush's handle allows for a firm grip.

Caption 4: The video captures a moment where a hand is applying a white substance, likely paint, to a dark surface using a toothbrush. The toothbrush is positioned diagonally across the frame, with the bristles facing the camera. The paint is being applied in a way that creates a textured effect, with some areas showing a smooth application and others showing a more uneven, brush-like texture. The hand is in the process of painting, with the fingers gripping the toothbrush and the thumb supporting it. The background is a solid, dark color, providing a stark contrast to the white paint, which enhances the visibility of the brush strokes. The overall layout is simple, focusing on the action of painting without any distractions.

Caption 5: The video captures a moment where a person is applying a white substance, likely paint, to a dark surface using a toothbrush. The action is deliberate and focused, with the bristles of the toothbrush making contact with the surface, creating a small, white splatter. The hand holding the toothbrush is steady, with fingers gripping the handle firmly. The background is a solid, dark color, providing a stark contrast to the white paint, which enhances the visibility of the splatter. The lighting is bright, casting a soft shadow of the toothbrush on the surface. The overall scene is static, with no movement or change in the environment, allowing the viewer to concentrate on the application process.

Caption 6: The video captures a moment where a person is applying a white substance, likely paint, to a dark surface using a toothbrush. The actions are deliberate and show a methodical approach, with the substance being carefully squeezed from a tube and then applied in a controlled manner. The agent's hand is steady, with fingers positioned to hold the tube and the toothbrush, indicating a focus on precision. The substance is being applied

in a way that suggests a deliberate effort to create a specific effect, possibly for artistic or decorative purposes. The overall scene is calm and focused, with the agent's hand being the main point of interest.

Caption 7: The video captures a moment where a toothbrush is being used to apply a white substance, likely a cleaning agent, to a dark surface. The substance appears to be in a liquid state, as evidenced by the way it drips and pools on the surface. The toothbrush is held in a way that suggests it is being used to spread the substance, with the bristles making contact with the surface. The fingers of the hand holding the toothbrush are visible, with one finger slightly raised, indicating the action of applying pressure. The background is a solid, dark color, providing a stark contrast to the white substance, which emphasizes the cleaning process.

Caption 8: The video captures a moment where a toothbrush is being used to apply a white substance, likely toothpaste, to a dark surface. The bristles of the toothbrush are in contact with the surface, and the substance is being squeezed out, creating a small puddle. The hand holding the toothbrush is steady, with fingers gripping the handle, indicating a controlled application. The surface appears smooth, with no visible resistance, allowing the substance to spread easily. The lighting is bright, casting a soft shadow of the toothbrush on the surface, enhancing the visibility of the application process.

Q1: Is the toothbrush handle described as plastic?
Q2: Did the paint cling to the bristles?
Q3: Was the paint stored in a can?
Q4: Was the canvas described as black?
Q5: Was the canvas backdrop described as vertical?
Q6: Did the person perform a flicking motion with the toothbrush?
Q7: Did droplets of paint detach from the bristles?
Q8: Was the detachment of paint caused by the force of the flick?
Q9: Did the paint droplets travel through the air?
Q10: Was gravity mentioned as influencing the droplets' travel?
Q11: Did the paint droplets adhere to the canvas surface?
Q12: Was the person's wrist motion described as controlled?
Q13: Did the person dip the bristles into the paint?
Q14: Was the paint described as thick?
Q15: Did the flicking motion require energy input from the person?

Respond only as:
Q1: Yes
Q2: No
...

E A Taxonomy of Physically Grounded Failure Modes

To analyze *why* certain prompts challenge current T2V models, we constructed a structured annotation framework that captures the physical-reasoning demands underlying each prompt, beyond surface interaction type. Every prompt is characterized along three complementary axes:

Action Complexity (AC).
- **AC-1 (Direct):** A single, simple action.
- **AC-2 (Goal-Oriented):** A preparatory or two-step action aimed at achieving a specific result.
- **AC-3 (Procedural):** A multi-step, complex procedure where correct sequencing is essential.

Scientific Principle (SP).
- **SP-1 (Basic):** Straightforward mechanics with no salient principle.
- **SP-2 (Applied):** Application of a common, understandable physical principle (e.g., friction, leverage).
- **SP-3 (Complex/Hidden):** Relies on less obvious or complex, multi-stage principles (e.g., negative pressure, magnetism).

Object Functionality (OF).
- **OF-1 (Intended Use):** Object used for its designed purpose.
- **OF-2 (Repurposed):** Object employed as an improvised tool.
- **OF-3 (System Component):** Object functioning within a broader physical system.

Illustrative Examples.
- **Example 1 (AC-1, SP-1, OF-1):**
  "Hands pour a handful of mixed coins into the opening of an empty, clear plastic milk jug."
  $\rightarrow$ *Direct transfer using object as intended; straightforward use of objects.*

- **Example 2 (AC-1, SP-1, OF-2):**
  "A plastic bucket is placed upside down on someone's head like a hat."
  → *Simple physical manipulation but repurposed functionality.*
- **Example 3 (AC-1, SP-2, OF-3):**
  "A lit flashlight inside a translucent plastic bag makes the bag glow like a lantern."
  → *Requires reasoning about light diffusion and emergent system behavior.*
- **Example 4 (AC-2, SP-3, OF-1):**
  "A person melts a toy soldier with a magnifying glass focusing sunlight."
  → *Goal-driven use of a non-obvious scientific principle (solar heat focus).*
- **Example 5 (AC-3, SP-1, OF-1):**
  "Hands pour colorful powder into a small cup, press it down firmly, then invert the cup and tap it, releasing a shaped bath bomb onto a surface."
  → *Multi-step procedural action with material handling.*
- **Example 6 (AC-3, SP-3, OF-3):**
  "Hands squeeze an empty plastic water bottle, place the opening over an egg yolk, and release the squeeze, sucking the yolk into the bottle."
  → *Complex, multi-step use of pressure dynamics and system-level interaction.*

To apply this taxonomy at scale, we used Gemini 2.5 Pro to classify every prompt in our Medium, Hard, and Very Hard subsets. As shown in Table A.2, the average scores increase consistently across all three axes, with harder groups involving more complex procedures, physical principles, and object functionality.

Table A.2: **Average taxonomy scores across model-based difficulty groups.** AC, SP, and OF denote Action Complexity, Scientific Principle, and Object Functionality, respectively. Each measure increases consistently from Medium to Hard to Very Hard, showing that the performance-based grouping also corresponds to progressively richer and more demanding physical-reasoning structure.

Difficulty Subset	AC	SP	OF	Overall Score
Medium Prompts	1.48	1.65	1.62	4.75
Hard Prompts	1.69	1.77	1.90	5.36
Very Hard Prompts	2.06	1.94	2.10	6.10

This structured analysis provides fine-grained diagnostic clarity: models often succeed on simple interactions (AC-1, SP-1, OF-1) but fail when tasks require multi-step procedures, non-obvious physical principles, or functional recontextualization. This taxonomy thus enhances PhysVidBench's value as a diagnostic tool, revealing not only how well models perform, but why they fail.

F An End-to-End Example Evaluation

Prompt Generation

Stage 1. Input: Human Commonsense Goal + Solutions (PIQA)

Goal: *Bag*
Solution: *can be used as a lantern with a flashlight indoors.*

Stage 2. Base Prompt:

A lit flashlight inside a translucent plastic bag makes the bag glow like a lantern.

Stage 3. Upsampled Prompt:

A cylindrical flashlight, currently switched on and emitting a focused beam of bright light, is positioned entirely inside a thin, flexible, translucent plastic bag. The lightweight plastic material loosely conforms around the flashlight. Light rays emanating from the flashlight's bulb strike the inner walls of the bag; due to the plastic's translucent property, the light scatters and diffuses as it passes through, causing the entire volume of the bag to glow softly from within, resembling a makeshift lantern in a dimly lit space.

Evaluation

Stage 1. Physics-Grounded Question Generation:

We generate questions that cover seven key physical commonsense dimensions defined in our ontology:

AU: Action & Procedural Understanding, **FM:** Force and Motion, **FP:** Fundamental Physics, **MT:** Material Interaction & Transformation, **OP:** Object Properties & Affordances, **SR:** Spatial Reasoning, **TD:** Temporal Dynamics.

#	Question	Dimension(s)
Q1	Is the flashlight emitting light energy?	FP
Q2	Is the bag made of plastic?	OP
Q3	Is the flashlight entirely inside the bag?	SR, OP
Q4	Is the bag described as flexible?	OP
Q5	Does the light scatter as it passes through the plastic?	MT, FP
Q6	Does the flashlight being on cause the bag to glow?	FP
Q7	Is the flashlight described as cylindrical?	SR, OP
Q8	Does the bag glow softly?	FP
Q9	Is the bag described as lightweight?	OP
Q10	Does the setup resemble a makeshift lantern?	AU
Q11	Is the light emitted by the flashlight described as a focused beam initially?	FP, OP
Q12	Does the entire volume of the bag glow?	SR, FP
Q13	Does the bag diffuse the light passing through it?	MT, FP

Stage 2. Dense Video Captioning:

The generated video is captioned using eight captions: one general-purpose caption and seven dimension-specific captions:

General:
The video features a close-up view of a bright light source, which appears to be a flashlight or a similar handheld light. The light is emitting a strong, white beam that illuminates the surrounding area with a soft glow. The light is focused and directed, creating a halo effect that enhances the visibility of the surrounding environment. The background is blurred, but hints of blue and white colors suggest a cool, possibly outdoor setting. The light source is partially obscured by a translucent, plastic bag, which adds a layer of texture to the scene. The overall composition is centered on the light, drawing attention to its brightness and the way it interacts with the environment

Fundamental Physics (FP):
The video captures a moment where a light source, likely a bulb, is enveloped in a translucent material, creating a visual effect that resembles a flower. The energy transfer is evident as the light shines through the material, creating a soft glow that illuminates the surrounding area. The material appears to be in a state of tension, with the light source at the center, creating a dynamic interplay of light and shadow. The balance is slightly tilted, with the light source being the focal point, drawing the viewer's attention. The physical state of the material is fluid, allowing the light to pass through, which changes the overall appearance of the object from a solid to a more ethereal, glowing form.

Object Properties & Affordances (OP):
The main objects in the video are the plastic bag and the light bulb. The plastic bag, with its smooth and reflective surface, allows light to pass through, creating a soft glow. The light bulb, with its metallic base, emits a bright light that illuminates the bag. The actions that can be directly observed are the light bulb emitting light and the plastic bag reflecting and diffusing the light. The materials of the bag and bulb are visually confirmed as soft and metallic, respectively. The bulb's rigidity is suggested by its shape and the way it stands upright.

Spatial Reasoning (SR):
The video captures a close-up view of a light bulb, which is the main subject. The bulb is encased in a clear, plastic bag, creating a soft glow that illuminates the surrounding area. The bag appears to be slightly crumpled, adding texture to the visual. The light emanating from the bulb creates a bright, circular area that contrasts with the darker background. The bulb is positioned centrally in the frame, with the plastic bag slightly tilted, allowing the light to shine outward. The overall layout is simple, focusing on the interplay of light and shadow.

Temporal Dynamics (TD):
The video captures a moment where a bright light source, likely a flashlight, is being directed towards a translucent, reflective surface, creating a dynamic interplay of light and shadow. The light appears to be moving, as evidenced by the varying intensity of the illumination and the shifting patterns of light and dark. The surface reflects the light, creating a soft glow that enhances the visual effect. The overall atmosphere is ethereal and slightly blurred, with the focus primarily on the light source and its reflection. The light source is the main subject, with the surrounding environment remaining indistinct, allowing the viewer to concentrate on the interplay of light and shadow.

Action & Procedural Understanding (AU):
The video captures a moment where a bright light source, possibly a flashlight, is being

directed towards a translucent, reflective surface, creating a dynamic interplay of light and shadow. The light source is the main focus, with its beam creating a luminous effect that illuminates the surrounding area. The reflective surface enhances the visual impact, as it captures and distorts the light, creating a sense of depth and movement. The agent's actions are deliberate, as they manipulate the light source to achieve the desired effect. The overall goal appears to be the exploration of light and reflection, showcasing the interplay between the light source and the reflective surface.

Material Interaction & Transformation (MT):
The video captures a moment where a clear plastic bag is being manipulated, likely by a hand, as it is folded and twisted. The bag's transparency allows the viewer to see the light source within, creating a bright, radiant effect. The light source appears to be the main focus, with its glow illuminating the bag and the surrounding environment. The bag's material, while initially clear, shows signs of distortion, possibly due to the folding and twisting actions. The light creates a soft glow that reflects off the bag's surface, enhancing the visual appeal of the scene. The overall interaction between the bag and the light source is one of visual interest, with the bag's transparency and the light's brightness creating a striking contrast.

Force and Motion (FM):
The video captures a moment where a clear plastic bag is being manipulated, likely by a hand, as evidenced by the slight blur around the edges of the bag. The bag is being stretched and folded, creating a dynamic interplay of light and shadow. The bright light source creates a lens flare effect, which adds to the visual complexity. The bag's transparency allows for a view of the light source, which is bright and white, indicating a strong light source. The bag's surface reflects the light, creating a shimmering effect. The overall scene is one of gentle interaction, with the bag responding to the touch of the hand, creating a fluid motion that is both delicate and deliberate.

Stage 3. QA-Based Evaluation (LLM-as-a-Judge)

#	Question	Answer
Q1	Is the flashlight emitting light energy?	Yes
Q2	Is the bag made of plastic?	Yes
Q3	Is the flashlight entirely inside the bag?	Yes
Q4	Is the bag described as flexible?	Yes
Q5	Does the light scatter as it passes through the plastic?	Yes
Q6	Does the flashlight being on cause the bag to glow?	Yes
Q7	Is the flashlight described as cylindrical?	No
Q8	Does the bag glow softly?	Yes
Q9	Is the bag described as lightweight?	No
Q10	Does the setup resemble a makeshift lantern?	No
Q11	Is the light emitted by the flashlight described as a focused beam initially?	Yes
Q12	Does the entire volume of the bag glow?	No
Q13	Does the bag diffuse the light passing through it?	Yes

G Additional Generation Results and Qualitative Examples

G.1 Qualitative Examples

To illustrate the behaviors and failure modes of current text-to-video models on PhysVidBench, we present several representative qualitative cases. For each case, we show: (1) the input text prompt, (2) key frames from videos generated by two contrasting models, and (3) a subset of yes/no QA outcomes used for scoring. *Note that only four questions are shown per example for brevity; the full evaluation uses a broader set of questions covering multiple physical-commonsense dimensions.* These examples highlight where models succeed or break down across different reasoning skills.



Prompt: Hands place a tea towel over a freshly baked cake with a domed top and gently press down, flattening the dome.

Prompt-based Evaluation Questions

Model: Cosmos (14B)

Q1: Does the fabric drape smoothly over the cake's dome?
A1: Yes

Q2: Is the tea towel described as rectangular?
A2: No

Q3: Does the cake visibly flatten as a result of the pressure?
A3: Yes

Q4: Does the warmth of the cake contribute to its structure yielding under pressure?
A4: No



Prompt: Hands mix baking soda and water in a small bowl to form a paste, then scoop some paste and rub it onto skin with orange streaks.

Prompt-based Evaluation Questions

Model: Cosmos (14B)

Q1: Did fingers mix the contents of the bowl?
A1: Yes

Q2: Was the resulting paste described as opaque?
A2: No

Q3: Were there orange streaks on the surface where the paste was applied?
A3: No

Q4: Was the bowl used for containment of the ingredients?
A4: Yes



Prompt: Toothpicks are pushed through slices of bread, securing them against the cut edges of a leftover cake.

Prompt-based Evaluation Questions

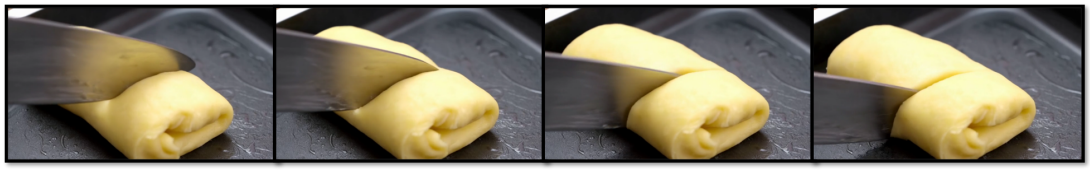
Model: Hunyuan

Q1: Are the toothpicks made of wood?
A1: Yes

Q2: Are the toothpicks held firmly?
A2: No

Q3: Do the toothpicks embed themselves within the cake's interior?
A3: Yes

Q4: Is the bread slice secured against the cake by the action?
A4: Yes



Prompt: The edge of a metal spoon presses down into a roll of soft potato dough, cutting off a small piece.

Prompt-based Evaluation Questions

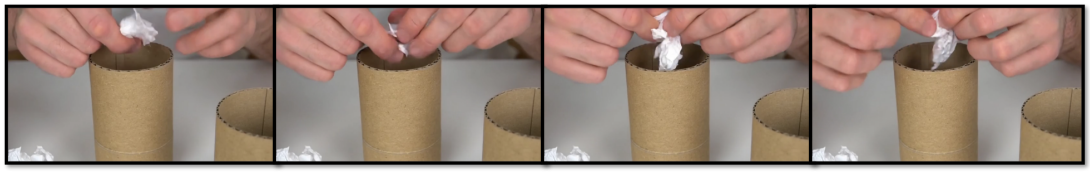
Model: Hunyuan

Q1: Does the action result in separating a portion of the dough?
A1: Yes

Q2: Is the dough described as having a cylindrical shape?
A2: Yes

Q3: Does the spoon shear through the dough's structure?
A3: No

Q4: Is the spoon described as being made of metal?
A4: No



Prompt: Hands stuff dryer lint into cardboard toilet paper tubes.

Prompt-based Evaluation Questions

Model: Wan2.1 (14B)

Q1: Is the tube held steady by one hand?
A1: No

Q2: Are two hands involved in the overall action described?
A2: Yes

Q3: Does the volume occupied by the lint decrease when pushed into the tube?
A3: Yes

Q4: Is force applied by fingers to push the lint?
A4: Yes



Prompt: Toothpaste is smeared onto the inner lens of swimming goggles and then wiped away with a soft cloth.

Prompt-based Evaluation Questions

Model: Wan2.1 (14B)

Q1: Did the toothpaste adhere to the lens surface after application?
A1: Yes

Q2: Was the cloth pressed against the lens?
A2: Yes

Q3: Did the wiping action remove the toothpaste from the lens?
A3: No

Q4: Is the goggle lens described as hard?
A4: No



Prompt: Hands place whole cherry tomatoes into the individual cups of a greased metal muffin tin.

Prompt-based Evaluation Questions

Model: LTX-Video

Q1: Is the muffin tin made of metal?
A1: Yes

Q2: Are the tomatoes described as small and round?
A2: No

Q3: Does gravity cause the released tomato to move downwards?
A3: No

Q4: Are the tomatoes placed into the cups one at a time?
A4: Yes



Prompt: Hands place several sheets of white styrofoam into the bottom of a terracotta planter.

Prompt-based Evaluation Questions

Model: LTX-Video

Q1: Do the hands lower the sheets into the planter?
A1: Yes

Q2: Do the sheets come to rest on the bottom surface of the planter?
A2: No

Q3: Is the styrofoam material described as porous?
A3: No

Q4: Is the planter described as sturdy?
A4: No



Prompt: A person attaches a binder clip to the end of a toothpaste tube and squeezes it to get more toothpaste out.

Prompt-based Evaluation Questions

Model: Wan2.1 (1.3B)

Q1: Does the clip's spring mechanism compress the tube?
A1: No

Q2: Is the toothpaste described as viscous?
A2: No

Q3: Does the binder clip create a barrier preventing toothpaste from moving back towards the flattened end?
A3: No

Q4: Does the clip help to gather the remaining toothpaste near the opening?
A4: No



Prompt: An ice cream scoop digs into cookie dough and releases a perfect ball onto a baking sheet.

Prompt-based Evaluation Questions

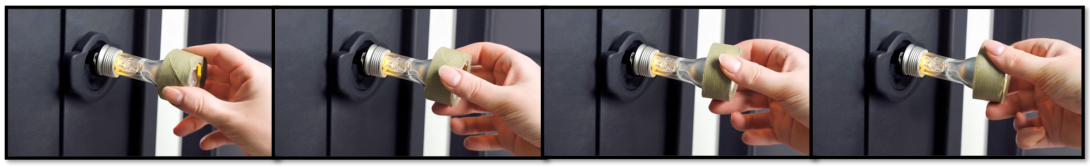
Model: Wan2.1 (1.3B)

Q1: Is the cookie dough described as soft?
A1: Yes

Q2: Does the dough yield under pressure from the scoop?
A2: No

Q3: Is the scoop positioned directly above the baking sheet before the dough is released?
A3: No

Q4: Is the ejected dough ball described as spherical?
A4: No



Prompt: Duct tape is wrapped around a stuck light bulb in a socket, and a hand uses the tape to twist the bulb counter-clockwise, loosening it.

Prompt-based Evaluation Questions

Model: Cosmos (7B)

Q1: Is the tape described as having a flexible fabric backing?
A1: No

Q2: Is the tape wrapped multiple times around the bulb?
A2: No

Q3: Does a hand grasp the layered duct tape on the bulb?
A3: No

Q4: Does the hand twist in a counter-clockwise direction?
A4: No



Prompt: Hands pour a handful of mixed coins into the opening of an empty, clear plastic milk jug.

Prompt-based Evaluation Questions

Model: Cosmos (7B)

Q1: Does the action of tilting the hands happen before the coins begin to fall?
A1: No

Q2: Does gravity influence the movement of the coins after release?
A2: No

Q3: Do the coins pass completely through the jug's opening to enter the body?
A3: No

Q4: Are the coins contained within the hands before being released?
A4: No



Prompt: Hands place folded dryer sheets inside a pair of sneakers.

Prompt-based Evaluation Questions

Model: VideoCrafter2

Q1: Do the hands push a dryer sheet into the first sneaker?
A1: No

Q2: Is the action of inserting a dryer sheet performed on both sneakers mentioned in the prompt?
A2: No

Q3: Is the dryer sheet described as flexible?
A3: No

Q4: Is the dryer sheet inserted into a cavity within the sneaker?
A4: No



Prompt: Hands place keys and folded money into a hollowed-out suntan lotion bottle and put the top back on.

Prompt-based Evaluation Questions

Model: VideoCrafter2

Q1: Are the keys described as rigid?
A1: No

Q2: Is the bottle described as hollowed-out?
A2: Yes

Q3: Are the keys inserted through an opening in the bottle?
A3: No

Q4: Is the bottle made of plastic?
A4: Yes



Prompt: Ice cubes are poured into a kitchen sink drain and ground up by the garbage disposal.

Prompt-based Evaluation Questions

Model: MAGI-1

Q1: Do the cubes tumble downwards into a cavity?
A1: No

Q2: Is the disposal chamber described as a confined space?
A2: No

Q3: Do the spinning components collide forcefully with the ice cubes?
A3: No

Q4: Do the internal components of the disposal spin rapidly?
A4: No



Prompt: A damp towel is spun rapidly in a room filled with visible smoke, causing the smoke to swirl and dissipate slightly.

Prompt-based Evaluation Questions

Model: MAGI-1

Q1: Was the air in the room initially still before the spinning?
A1: Yes

Q2: Does the damp surface of the towel interact with the smoke particles?
A2: No

Q3: Is the towel's motion described as being driven against air resistance?
A3: Yes

Q4: Does the rapid movement of the towel generate air currents?
A4: Yes



Prompt: A hand wipes a glass mirror clean using a crumpled brown paper bag.

Prompt-based Evaluation Questions

Model: CogVideoX (2B)

Q1: Is the paper bag described as flexible?
A1: Yes

Q2: Does the manipulation of the bag into a wad happen *before* it is pressed against the mirror?
A2: No

Q3: Is the described action a wiping motion?
A3: No

Q4: Is pressure applied by the hand onto the paper wad against the mirror?
A4: No



Prompt: A decorative automobile bucket seat cover is stretched and fitted over an old, worn desk chair.

Prompt-based Evaluation Questions

Model: CogVideoX (2B)

Q1: Was the cover pulled over the backrest first?
A1: No

Q2: Did the application process involve stretching the cover?
A2: Yes

Q3: Does the cover conform closely to the chair's shape?
A3: Yes

Q4: Does the chair have underlying padding mentioned?
A4: No



Prompt: Clear plastic sheeting is carefully draped over several small tomato plants in a garden bed at dusk.

Prompt-based Evaluation Questions

Model: CogVideoX (5B)

Q1: Does gravity cause the plastic sheet to move downwards?
A1: No

Q2: Does the plastic sheet change shape as it settles on the plants?
A2: Yes

Q3: Does the plastic sheet cover the tomato plants?
A3: Yes

Q4: Does the flexibility of the plastic allow it to bend over the plants?
A4: Yes



Prompt: Coarse salt is poured into a dirty cast iron skillet, then scrubbed vigorously with the cut side of a potato half.

Prompt-based Evaluation Questions

Model: CogVideoX (5B)

Q1: Does the hand apply pressure while rubbing the potato?
A1: No

Q2: Is the salt ground between the potato and the skillet surface?
A2: No

Q3: Is consistent pressure applied by the hand?
A3: No

Q4: Is the salt contained within the skillet's interior after pouring?
A4: No



Prompt: A shirt sleeve draped over the edge of a wooden table is being pressed flat by a moving iron.

Prompt-based Evaluation Questions

Model: Sora

Q1: Does the iron apply downward pressure onto the fabric?
A1: Yes

Q2: Does the iron transfer heat to the fabric?
A2: Yes

Q3: Is the soleplate of the iron made of metal?
A3: Yes

Q4: Is the soleplate of the iron flat?
A4: No



Prompt: Hands cut strands of yarn with scissors, then apply glue to the ends and press them onto a mannequin head.

Prompt-based Evaluation Questions

Model: Sora

Q1: Does squeezing the scissor handles cause the blades to pivot?
A1: Yes

Q2: Is the adhesive described as a liquid when applied?
A2: No

Q3: Are multiple strands of yarn held by the hands initially?
A3: No

Q4: Does the cutting action result in loose strands?
A4: Yes



Prompt: Toothpaste is squeezed onto a scratched CD, then rubbed radially outwards with a finger.

Prompt-based Evaluation Questions

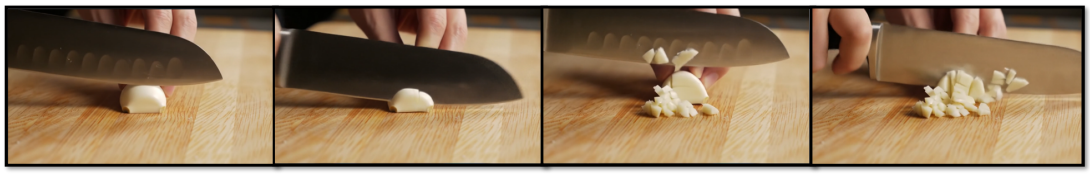
Model: Veo-2

Q1: Is the finger used to apply force to spread the paste?
A1: Yes

Q2: Is the CD's state of rest maintained throughout the toothpaste application and spreading?
A2: No

Q3: Does the rubbing action cause the toothpaste to cover the scratched area?
A3: No

Q4: Is the CD described as rigid?
A4: Yes



Prompt: The flat side of a chef's knife is pressed down onto a garlic clove on a cutting board, crushing it.

Prompt-based Evaluation Questions

Model: Veo-2

Q1: Does the garlic clove audibly crack?
A1: No

Q2: Is downward force applied to the knife?
A2: Yes

Q3: Is the knife described as rigid?
A3: Yes

Q4: Is the knife positioned parallel to the cutting board surface?
A4: No

G.2 Model Generation Specifications

Table A.3 summarizes, for each evaluated text-to-video (T2V) model, the output resolution, number of frames, video duration, and guidance scale used throughout our experiments. We use these settings consistently when generating both base and enriched videos to ensure a fair comparison across models.

Table A.3: **Model generation specifications.** This table lists the generation settings for each text-to-video (T2V) model used in our experiments. For each model, we report the video resolution (in width and height), number of frames, total video duration (in seconds), and the guidance scale used during sampling (if applicable). These settings reflect the default or recommended configurations for each model and were held consistent across evaluations to ensure comparability in visual output and downstream analysis.

Model	Resolution	#Frames	Duration	Guidance Scale
LTX-Video	1216×704	121	4	3
VideoCrafter2	512×320	16	1	12
CogVideoX (2B)	720×480	49	6	6
CogVideoX (5B)	720×480	49	6	6
Wan2.1 (1.3B)	832×480	33	6	5
Wan2.1 (14B)	832×480	33	6	5
MAGI-1	720×720	48	2	7.5
HunyuanVideo	960×544	129	5	6
Cosmos (7B)	1280×704	121	5	7
Cosmos (14B)	1280×704	121	5	7
Sora	1280×720	150	5	–
Veo-2	1280×720	120	5	–

G.3 Effect of Caption Detail on Model Performance: Short vs. Dense Captions

Caption detail strongly affects the accuracy and reliability of VLM-based evaluation. Short captions tend toward high-level summaries and omit subtle but critical details, e.g. precise object interactions or slight motions, which leads to incomplete or inaccurate physical commonsense judgments. When queried directly, VLMs may also hallucinate or exaggerate minor cues, further degrading evaluation quality. Dense captioning instead records fine-grained observations without suggestive prompting. We use AuroraCap for its ability to generate detailed, multi-faceted descriptions: for each video we produce one general-purpose caption and seven dimension-specific captions aligned with our ontology (e.g., spatial layout, force dynamics, material behavior). A yes/no question is marked correct if any of the eight pooled captions supply the required evidence. Using complementary views reduces dependence on any single caption and provides broader coverage of the visual evidence relevant to physical reasoning.

Sec. I.3 further evaluates caption recall and false-positive behavior.

H User Study Details and Human Evaluation

To verify that our automatic, caption-based evaluation (auto-eval pipeline) aligns with human judgment, we carried out a user study via Qualtrics. We selected the four top-performing open-source models, namely Cosmos-14B, Wan2.1-14B, MAGI-1, and HunyuanVideo, and sampled 30 prompts (out of 383 upsampled prompts) at random. Each prompt was used to generate one video per model, yielding 120 videos in total.

Procedure. For each video, we used the yes/no questions produced by our auto-eval pipeline and presented them in a survey interface built with Qualtrics that mimics the automated QA layout (Figure A.2). Participants saw each video alongside five randomly assigned questions, answered in a Yes/No format. Videos and questions were fully randomized to prevent any inference of prompt or model identity. Before beginning, participants read brief instructions on interpreting the questions; we then collected non-identifying metadata (age range and self-reported familiarity with deep learning).

Participants. Fifteen individuals took part, all

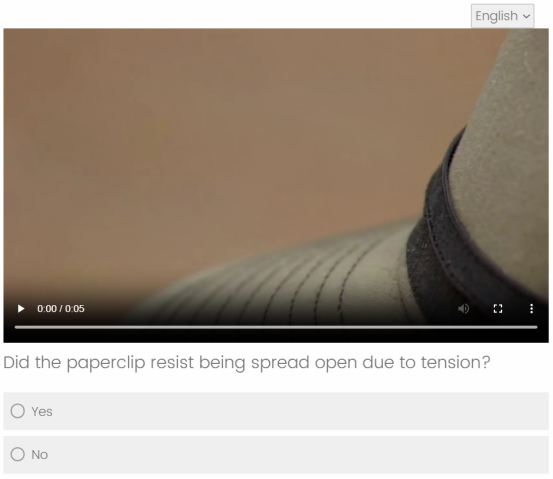


Figure A.2: **User Study Interface.** Our web-based interface for the human study mimicked the auto-eval QA structure. Participants watched a video, then answered five Yes/No questions drawn from the same QA bank used in automatic evaluation. The design enforced randomization across videos and questions, while masking model and prompt identity to reduce bias.

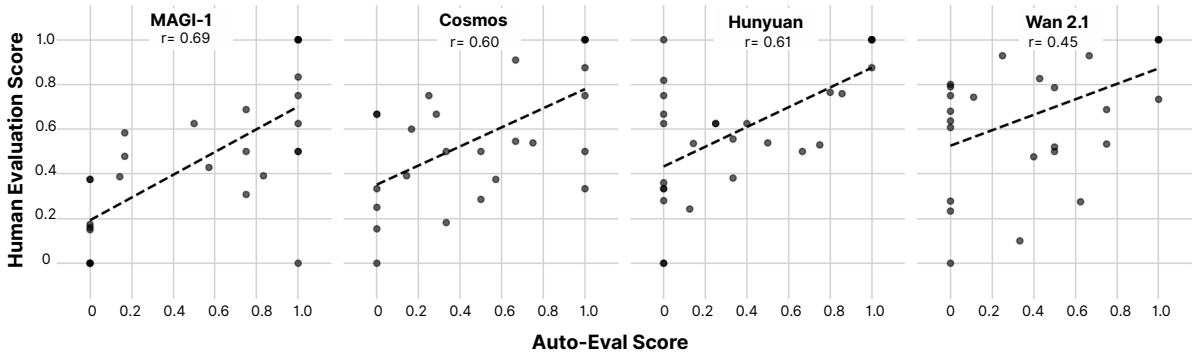


Figure A.3: **Pearson correlation between human and automatic evaluations.** Each subplot compares scores over 30 prompts and reports a linear fit for one open-source generator: MAGI-1 ($r = 0.69$), Cosmos-14B ($r = 0.60$), HunyuanVideo ($r = 0.61$), and Wan2.1-14B ($r = 0.45$).

volunteers, collectively providing 1,340 binary (Yes/No) responses across the 120 videos. Owing to the randomized assignment and modest pool size, some questions were answered only once, while others received multiple responses.

Results and Analysis. We computed Pearson correlation coefficients to measure alignment between auto-eval scores and human judgments. As expected, correlation rose with the number of annotators. For MAGI-1, questions answered by only a single participant gave r=0.47, while those answered by at least three reached r=0.69. Figure A.3 plots these correlations with linear fits for MAGI-1, Cosmos-14B, HunyuanVideo, and Wan2.1-14B. Although Wan2.1-14B attains the lowest model-level correlation (r=0.45, Table 5), this remains competitive given the task's ambiguity and binary scoring. VideoPhy-2 reports an average Pearson correlation of r=0.285 for the prior VideoCon-Physics evaluator introduced in VideoPhy and r=0.42 for its own AutoEval model on unseen prompts (Bansal et al., 2025, 2026), whereas our evaluator reaches r=0.69 at best in our human study. Together, these findings indicate that the caption→LLM design is a reliable proxy for human judgment, and that adding annotators strengthens the alignment further. Inter-annotator agreement is also consistent across generators. For items with at least three responses, Krippendorff's α is 0.74 for MAGI-1, 0.77 for Cosmos-14B, 0.71 for HunyuanVideo, and 0.74 for Wan2.1-14B; with at least four responses, the corresponding values are 0.80, 0.85, 0.75, and 0.80.

I Robustness of the Automatic Evaluation Pipeline

We designed our automatic caption-based evaluation ("auto-eval") with robustness as a primary objective. This section compiles the empirical evidence for its reliability and the analyses we used to stress-test each component.

I.1 Cross–Judge Stability

To test whether results depend on a particular judge model, we replaced Gemini-Flash with DeepSeek. The induced ranking of T2V models is highly stable (Spearman $\rho = 0.98$ over the full model list), showing that the pipeline measures differences between videos rather than idiosyncrasies of the judge (Table A.4).

Table A.4: **Cross–judge stability of our evaluation pipeline.** Replacing Gemini–Flash with DeepSeek yields nearly identical model rankings, indicating the performance signal is not judge–specific but emerges from our language-grounded evaluation design.

Model	Gemini–Flash Rank	DeepSeek Rank
Cosmos (14B)	1	1
Cosmos (7B)	2	3
Wan2.1 (14B)	3	2
MAGI–1	4	4
Wan2.1 (1.3B)	5	5
HunyuanVideo	6	6
CogVideoX (2B)	7	7
VideoCrafter2	8	9
CogVideoX (5B)	9	8
LTX–Video	10	10
Spearman $\rho = 0.98$ (rank stability)		

I.2 Captioner Analysis and Cross–Captioner Stability

To assess whether PhysVidBench depends on a particular captioning model, we re-ran the full pipeline with two alternative captioners, Qwen2.5-VL-72B-Instruct-AWQ and Qwen3-VL-30B-A3B-Instruct, in place of AuroraCap. Agreement is strong across all ten open-source T2V models: Spearman rank correlations between captioners range from 0.80 to 1.00, so the relative ordering of models is effectively unchanged. Cosmos-14B, Wan2.1-14B, MAGI-1, and HunyuanVideo remain the top-performing models under every captioning variant, indicating that the evaluator captures properties of the videos rather than of any single captioner. Pearson correlations of 0.74–0.98 further confirm that absolute scores move coherently across captioners. Table A.5 summarizes these results.

Table A.5: **Cross-captioner agreement.** Lower triangle: Spearman rank correlation; upper triangle: Pearson correlation, over model-level accuracies. Agreement is high between the two Qwen captioners and moderate-to-high for AuroraCap, indicating PhysVidBench scores are largely stable across captioners.

	AuroraCap	Qwen-2.5-VL	Qwen-3-VL
AuroraCap	–	0.83	0.74
Qwen-2.5-VL	0.80	–	0.98
Qwen-3-VL	0.80	1.00	–

These results show that PhysVidBench is not captioner-dependent: swapping in significantly different captioners, including models with different vision–language backbones and training corpora, yields essentially the same model rankings and similar absolute scores. Together with the multi-caption design, this consistency indicates that the evaluation is robust to the choice of captioner.

I.3 Caption Evidence Recall and Mismatched-Video Control

We first measure caption-stage recall on the 120 videos used in our human study. Two expert annotators identify the physically relevant events visible in each video and check whether each event appears in at least one of the eight generated captions. An event is counted as a false negative only when it is visibly present but absent from every caption. This analysis yields 98.5% event-level recall, corresponding to a 1.5% false-negative rate.

We then test the complementary false-positive pathway by pairing each question set with a video generated by the same model from a different, randomly selected prompt. We apply this control to all 4,123 questions for every evaluated model. As shown in Table A.6, positive-response rates remain between 2.8% and 5.1% across all models and physical-commonsense dimensions, substantially below the matched accuracies in Table 3. This consistent collapse indicates that the evaluator relies on evidence from the corresponding video rather than question wording or a general tendency to answer "Yes."

Because unrelated videos may occasionally satisfy individual questions by chance, these residual positive responses provide a conservative estimate of the evaluator's false-positive tendency. Together, the high caption recall and low mismatched-video response rates show that relevant evidence is rarely omitted, while unrelated visual evidence is rarely accepted.

I.4 Error-Guided Prompt Refinement

Beyond passive scoring, we used auto–eval as a diagnostic signal in a training–free self–refinement loop: detect failed dimensions $\rightarrow$ prompt rewrite $\rightarrow$ regenerate $\rightarrow$ re–evaluate. On 100 prompts, CogVideoX-2B and CogVideoX-5B improved from 21.6$\rightarrow$32.7 (+11.1 percentage points) and 17.8$\rightarrow$29.7 (+11.9 percentage points), respectively. This iterative experiment is separate from one-shot prompt enrichment and was evaluated only on these two models.

An Illustrative Example

The example below shows how PhysVidBench feedback guided a revised prompt that led to a visibly more plausible generation:

Base Prompt: *A screwdriver tip is placed on the lid of a metal can, and a hammer taps the screwdriver handle, puncturing the lid.*
Refined Prompt (Guided by Feedback): *A person grips a hammer, delivering a sharp, controlled tap to the rigid upper end of a vertically positioned screwdriver with a solid shaft; the heavy hammer's impact force travels down the screwdriver's shaft, concentrating at its tip, which is pressed firmly against the surface of a sealed metal can lid, causing the thin metal to yield and puncture after the hammer tap, overcoming the lid's structural integrity.*

Table A.6: Mismatched-video control by physical commonsense dimension. Each question set is paired with a video generated by the corresponding model from a different prompt. Values report positive-response rates (%). AU = Action & Procedural Understanding, FM = Force and Motion, FP = Fundamental Physics, MT = Material Interaction & Transformation, OP = Object Properties & Affordances, SR = Spatial Reasoning, and TD = Temporal Dynamics.

Model	AU	FM	FP	MT	OP	SR	TD	Average
LTX-Video	3.4	3.7	3.1	4.0	4.5	3.3	2.9	3.6
VideoCrafter2	3.2	3.5	3.8	3.6	4.2	3.0	2.8	3.4
CogVideoX (2B)	3.8	4.0	3.6	4.1	4.7	3.5	3.1	3.8
CogVideoX (5B)	3.5	3.8	3.4	3.9	4.4	3.2	3.0	3.6
Wan2.1 (1.3B)	4.0	4.3	3.9	4.4	4.9	3.8	3.4	4.1
Wan2.1 (14B)	4.2	4.5	4.1	4.6	5.1	4.0	3.6	4.3
MAGI-1	3.9	4.2	3.8	4.3	4.8	3.7	3.3	4.0
HunyuanVideo	3.7	4.0	3.6	4.2	4.6	3.5	3.2	3.8
Cosmos (7B)	4.1	4.4	4.0	4.5	5.0	3.9	3.5	4.2
Cosmos (14B)	4.3	4.6	4.2	4.7	5.1	4.1	3.7	4.4
Sora	3.8	4.1	3.7	4.3	4.8	3.6	3.2	3.9
Veo-2	4.0	4.3	3.9	4.4	4.9	3.8	3.4	4.1

Taken together, these analyses show that our evaluation pipeline (i) aligns more strongly with human judgments than the other evaluated automatic methods, (ii) yields stable rankings across different judges and captioners, (iii) achieves high caption-event recall on the annotated human-study videos, (iv) shows a sharp drop in positive responses under the mismatched-video control for Wan2.1-14B, and (v) can provide diagnostic feedback for prompt refinement. Together, these results support the robustness of the caption-grounded evaluation pipeline for physical commonsense video generation.

For dimension-level uncertainty, we perform 10,000 prompt-level bootstrap resamples. Each sampled prompt retains all of its associated QA items, preserving dependence among questions derived from the same prompt. Table A.7 reports the resulting 95% interval half-widths.

I.5 Statistical Significance and Confidence Interval Analysis

Bootstrap confidence intervals provide evidence that PhysVidBench yields statistically interpretable evaluation signals. Using 10K resamples, we find that the confidence interval on the enriched-minus-base accuracy difference lies entirely above zero for six of ten models. For CogVideoX-5B, Wan2.1-14B, VideoCrafter2, and HunyuanVideo, the interval includes zero. Overall, prompt enrichment yields modest, model-dependent gains, with statistically supported improvements for six of the ten models.

For the model-based difficulty groups, the Medium, Hard, and Very Hard intervals do not overlap for any model under either base or upsampled prompts (Tables A.9 and A.10), showing a consistent separation in model performance across the three groups. Together with the taxonomy analysis in Table A.2, where action complexity, scientific-principle complexity, and object functionality increase monotonically from Medium to Very Hard, this indicates that the performance-based grouping aligns with increasingly demanding physical-reasoning characteristics rather than arbitrary score partitions.

Overall, the bootstrap analysis quantifies uncertainty in the reported results, supports a consistent distinction across the model-based difficulty groups, and shows that the effect of prompt enrichment is modest and model-dependent.

Acknowledgements

We thank all reviewers for their valuable comments and suggestions, which helped improve the final version of this work. This research was supported in part by the KUIS AI Center Research Awards to Enes Sanli. Aykut Erdem acknowledges funding from the TÜBİTAK 2247-A Program under Award 123C550. The authors also gratefully acknowledge the use of Google Cloud Platform credits provided through Google's Gemini Academic Program Award. Generative AI assistants were used for language polishing, paraphrasing, and drafting assistance. All scientific content, analyses, claims, citations, and final text were reviewed and verified by the authors.

Table A.7: **Prompt-level bootstrap 95% confidence intervals by physical dimension.** Values are upsampled-prompt caption-grounded QA accuracy percentages ± the 95% bootstrap interval half-width. AU = Action & Procedural Understanding, FM = Force and Motion, FP = Fundamental Physics, MT = Material Interaction & Transformation, OP = Object Properties & Affordances, SR = Spatial Reasoning, and TD = Temporal Dynamics. Prompts, rather than individual QA items, are resampled.

Model	AU	FM	FP	MT	OP	SR	TD
LTX-Video	20.7 ± 3.40	19.9 ± 3.10	18.6 ± 3.15	16.9 ± 3.75	25.0 ± 2.95	17.3 ± 2.80	13.7 ± 4.35
VideoCrafter2	17.7 ± 2.90	19.4 ± 2.95	23.3 ± 3.75	18.8 ± 4.15	26.6 ± 2.75	15.8 ± 2.55	12.6 ± 4.05
CogVideoX (2B)	22.6 ± 3.15	24.6 ± 2.95	23.8 ± 3.70	22.6 ± 4.10	28.9 ± 2.75	19.6 ± 2.85	16.8 ± 4.60
CogVideoX (5B)	18.9 ± 3.00	19.6 ± 2.95	21.3 ± 3.40	18.1 ± 3.90	25.6 ± 2.75	17.9 ± 2.90	12.6 ± 4.00
Wan2.1 (1.3B)	30.1 ± 3.55	29.0 ± 3.30	28.5 ± 4.00	28.9 ± 4.60	35.2 ± 2.95	27.4 ± 3.25	18.7 ± 4.85
Wan2.1 (14B)	33.6 ± 3.65	31.6 ± 3.50	32.6 ± 4.35	32.9 ± 4.80	39.0 ± 2.95	29.6 ± 3.30	23.7 ± 5.20
MAGI-1	27.3 ± 3.35	26.6 ± 3.25	29.4 ± 4.20	28.5 ± 4.45	37.4 ± 3.00	30.4 ± 3.30	19.1 ± 4.85
HunyuanVideo	26.7 ± 3.45	26.2 ± 3.15	28.1 ± 3.95	32.3 ± 4.75	36.1 ± 3.00	23.9 ± 3.10	21.8 ± 5.25
Cosmos (7B)	32.5 ± 3.85	32.8 ± 3.70	33.6 ± 4.15	36.1 ± 4.90	40.1 ± 2.90	31.6 ± 3.55	21.4 ± 5.30
Cosmos (14B)	35.6 ± 3.50	33.3 ± 3.35	31.7 ± 3.90	37.4 ± 4.60	40.9 ± 2.90	32.5 ± 3.25	21.0 ± 5.05
Sora	28.6 ± 3.20	30.0 ± 3.25	29.9 ± 4.20	33.1 ± 4.80	37.0 ± 2.60	24.8 ± 2.90	16.7 ± 4.55
Veo-2	34.9 ± 3.35	33.8 ± 3.40	31.7 ± 4.20	36.0 ± 4.55	38.4 ± 2.85	29.7 ± 3.30	19.5 ± 5.85

Table A.8: **Bootstrap 95% confidence intervals for model accuracies using upsampled prompts, base prompts, and their paired differences.** Positive intervals in the "Difference" column indicate statistically significant improvements from prompt enrichment.

Model	Upsampled Prompts	Base Prompts	Difference
CogVideoX (2B)	[0.2341, 0.2784]	[0.2042, 0.2458]	[0.0068, 0.0556]
CogVideoX (5B)	[0.1883, 0.2317]	[0.1735, 0.2162]	[-0.0097, 0.0399]
Wan2.1 (1.3B)	[0.2831, 0.3309]	[0.2525, 0.2977]	[0.0075, 0.0561]
Wan2.1 (14B)	[0.3165, 0.3658]	[0.2982, 0.3468]	[-0.0078, 0.0448]
VideoCrafter2	[0.2012, 0.2405]	[0.1832, 0.2237]	[-0.0013, 0.0355]
LTX-Video	[0.1883, 0.2325]	[0.1604, 0.2009]	[0.0071, 0.0526]
Cosmos (7B)	[0.3248, 0.3762]	[0.2436, 0.2910]	[0.0555, 0.1109]
Cosmos (14B)	[0.3398, 0.3847]	[0.2712, 0.3117]	[0.0463, 0.0956]
HunyuanVideo	[0.2769, 0.3248]	[0.2675, 0.3153]	[-0.0168, 0.0352]
MAGI-1	[0.3013, 0.3505]	[0.2489, 0.2961]	[0.0276, 0.0795]

Table A.9: **Bootstrap 95% confidence intervals across model-based difficulty groups for upsampled prompts.** Each interval reflects model accuracy uncertainty within the Very Hard, Hard, and Medium groups. The non-overlapping intervals show a consistent separation in model performance across the three groups. Together with the taxonomy analysis in Table A.2, this separation is associated with increasing action complexity, scientific-principle complexity, and object functionality from Medium to Very Hard.

Model	Very Hard Prompts	Hard Prompts	Medium Prompts
CogVideoX (2B)	[0.0961, 0.1365]	[0.2551, 0.3087]	[0.4429, 0.5555]
CogVideoX (5B)	[0.0897, 0.1382]	[0.1934, 0.2481]	[0.3248, 0.4596]
Wan2.1 (1.3B)	[0.1227, 0.1697]	[0.3103, 0.3663]	[0.5210, 0.6285]
Wan2.1 (14B)	[0.1547, 0.2060]	[0.3419, 0.3998]	[0.5645, 0.6642]
VideoCrafter2	[0.0876, 0.1275]	[0.2074, 0.2553]	[0.3958, 0.4938]
LTX-Video	[0.0729, 0.1110]	[0.2009, 0.2573]	[0.3489, 0.4742]
Cosmos (7B)	[0.1552, 0.2182]	[0.3574, 0.4186]	[0.5488, 0.6606]
Cosmos (14B)	[0.1951, 0.2487]	[0.3651, 0.4201]	[0.5528, 0.6319]
HunyuanVideo	[0.1336, 0.1859]	[0.3056, 0.3646]	[0.4667, 0.5830]
MAGI-1	[0.1447, 0.1911]	[0.3229, 0.3830]	[0.5483, 0.6589]

Table A.10: **Bootstrap 95% confidence intervals across model-based difficulty groups for base prompts.** The non-overlapping intervals across Medium, Hard, and Very Hard groups show that the same separation in model performance persists for base prompts, indicating that the observed difficulty structure is not driven by prompt enrichment.

Model	**Very Hard Prompts**	**Hard Prompts**	**Medium Prompts**
CogVideoX (2B)	[0.0901, 0.1292]	[0.2251, 0.2775]	[0.3619, 0.4806]
CogVideoX (5B)	[0.0710, 0.1151]	[0.1856, 0.2419]	[0.3026, 0.4287]
Wan2.1 (1.3B)	[0.1110, 0.1577]	[0.2650, 0.3189]	[0.4849, 0.5849]
Wan2.1 (14B)	[0.1543, 0.2127]	[0.3140, 0.3729]	[0.5065, 0.6238]
VideoCrafter2	[0.0729, 0.1105]	[0.1949, 0.2448]	[0.3501, 0.4544]
LTX-Video	[0.0604, 0.0979]	[0.1601, 0.2083]	[0.3358, 0.4560]
Cosmos (7B)	[0.1129, 0.1602]	[0.2556, 0.3155]	[0.4314, 0.5623]
Cosmos (14B)	[0.1689, 0.2254]	[0.2810, 0.3303]	[0.4147, 0.5076]
HunyuanVideo	[0.1108, 0.1558]	[0.2886, 0.3493]	[0.5129, 0.6097]
MAGI-1	[0.1126, 0.1621]	[0.2675, 0.3264]	[0.4401, 0.5545]