VIDEOPHY-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation

Hritik Bansal[*1] **Clark Peng**[*1] **Yonatan Bitton**[*2]
Roman Goldenberg[2] **Aditya Grover**[1] **Kai-Wei Chang**[1]

[1]**University of California Los Angeles** [2]**Google Research**

Abstract

Large-scale video generative models, capable of creating realistic videos of diverse visual concepts, are strong candidates for general-purpose physical world simulators. However, their adherence to physical commonsense across real-world actions remains unclear (e.g., playing tennis, backflip). Existing benchmarks suffer from limitations such as limited size, lack of human evaluation, sim-to-real gaps, and absence of fine-grained physical rule analysis. To address this, we introduce VIDEOPHY-2, an action-centric dataset for evaluating physical commonsense in generated videos. We curate 200 diverse actions and detailed prompts for video synthesis from modern generative models. We perform human evaluation that assesses semantic adherence, physical commonsense, and grounding of physical rules in the generated videos. Our findings reveal major shortcomings, with even the best model achieving only 22% joint performance (i.e., high semantic and physical commonsense adherence) on the hard subset of VIDEOPHY-2. We find that the models particularly struggle with conservation laws like mass and momentum. Finally, we also train VIDEOPHY-2-AUTOEVAL, an automatic evaluator for fast, reliable assessment on our dataset. Overall, VIDEOPHY-2 serves as a rigorous benchmark, exposing critical gaps in video generative models and guiding future research in physically-grounded video generation. The data and code is available at https://videophy2.github.io/.

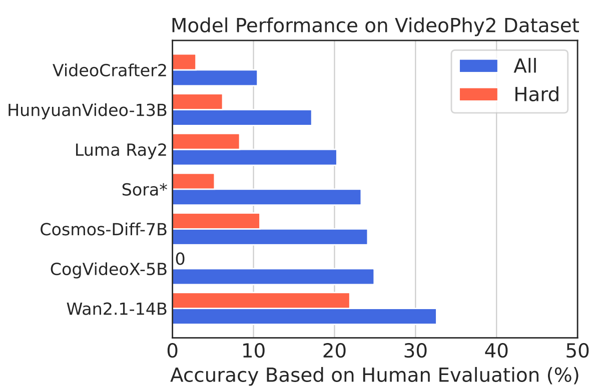
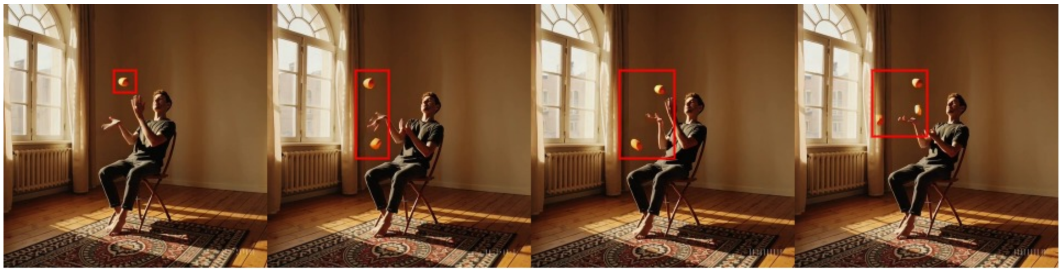
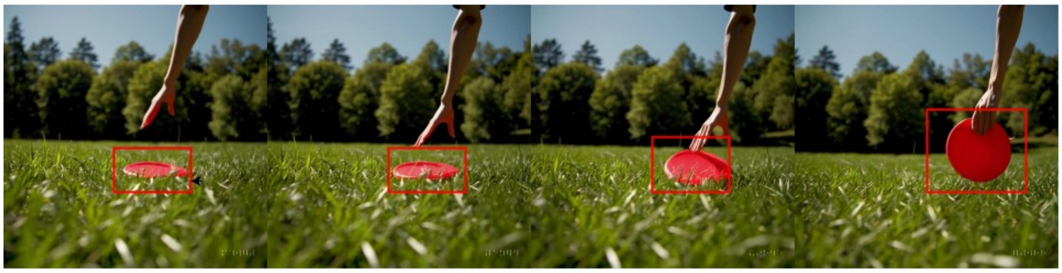


Figure 1: **Performance on the VIDEOPHY-2 dataset using human evaluation.** We evaluate the physical commonsense and semantic adherence to text conditioning prompts for diverse real-world actions. We observe that even the best-performing model Wan2.1-14B achieves 32.6% and 22% on the entire and hard subset of the data, respectively. * represents the evaluation on a very small subset of the dataset.

[*] Equal Contribution.

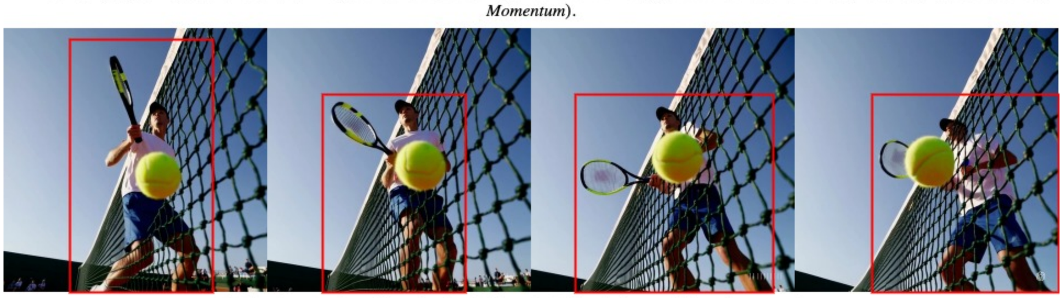


(a) **Caption**: A person sits and juggles three small oranges, using their legs to help with the pattern.
Human-judged Physical Violations: The number of juggling balls must stay constant and follow a parabolic downward trajectory (*Gravity, Conservation of Mass*).



(b) **Caption**: A person throws a frisbee that tumbles along the ground before stopping.
Human-judged Physical Violations: The frisbee must contact the hand before any upward movement occurs (*Conservation of Momentum*).

(c) **Caption**: A clay pot falls from a shelf and hits the tiled floor, breaking into pieces, and coming to rest.
Human-judged Physical Violations: The pot should also break upon contact with the ground and the other shards (*Hardness, Conservation of Momentum*).



(d) **Caption**: A player hits a forehand drive, the ball spinning rapidly as it crosses the net.
Human-judged Physical Violations: The player and racket must not pass through the net during play (*Hardness, Friction*).

(e) **Caption**: Multiple candles of varying heights and widths are blown out simultaneously by a single breath, some flames extinguishing faster than others.
Human-judged Physical Violations: If the candles are lit and unwavering, they should emit a constant source of light (*Combustion*).

Figure 2: **Examples of physically unlikely video generations from Sora.** Each case demonstrates violations of physical rules and their laws.

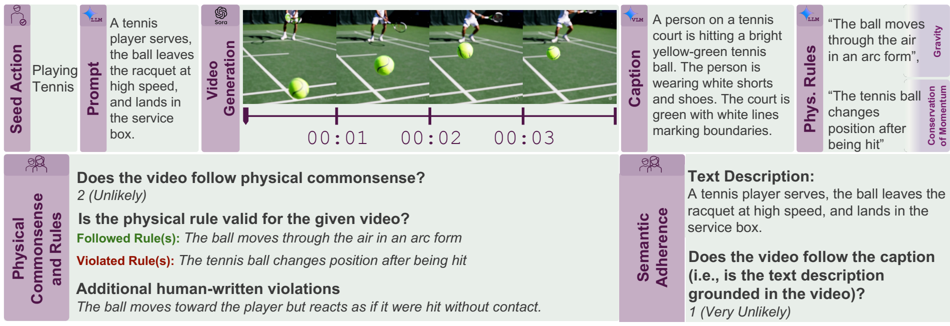


Figure 3: **VIDEOPHY-2 pipeline. Top:** We generate a text prompt from the seed action using an LLM, create a video with a text-to-video model, and caption it with a VLM to extract candidate physical rules. **Bottom:** Human annotators rate the video's physical likelihood, verify rule violations, suggest missing rules, and assess semantic adherence to the input prompt.

1 Introduction

Recent advancements in large-scale video generative modeling offer the potential to simulate the physical world accurately [10, 53]. In particular, this capability can enable learning general-purpose visuomotor policies [33, 16], predicting future frames for robotic manipulation [1], autonomous driving [1], and game playing [12, 18]. In daily life, humans rely on their sophisticated physics intuition to interact with the world [17] (e.g., predicting the trajectory of football after being hit). However, the extent to which existing video models can generate physically likely worlds across diverse real-world actions remains unclear.

A naive approach to evaluating generated videos is to compare them with ground-truth physical simulations [1, 47]. Furthermore, there is a lack of mature methods for rendering diverse real-world materials [5, 28, 45] and for accurately simulating complex physical interactions [36]. For instance, simulating a scenario like 'a child kicking a ball against a wall' requires precise estimation of the foot's pose and geometry relative to the ball at impact, an exact model of the kick's dynamics, and considerations of the ball's air pressure and material properties. While we focus on evaluating the physical likelihood of generated videos, an assessment that can often be made by humans without formal physics education by relying on their real-world experience.

Recent work such as Physics-IQ [44] conditions video models on the first few frames of real videos and evaluates their similarity by comparing predicted videos with ground-truth completions. However, this approach faces several challenges: (a) the extent to which it agrees with human judgment remains unclear, and (b) extending it to more complex scenarios depicting multiple events is non-trivial. Another work PhyGenBench [36] curates a small set of 160 manually crafted prompts, which is not scalable. Additionally, their evaluation approach simplifies the problem by designing text prompts that explicitly associate with a single physical law (e.g., 'A stone placed on the surface of a water pool' is linked to law of Buoyancy). Although, this strict one-to-one association between a prompt and a physical law is problematic, as video models often exhibit imperfect semantic adherence. For instance, a video model might generate a video that does not strictly follow the prompt but still adheres to physical commonsense (e.g., producing a video where 'a stone is dropped from a height into the pool', where gravity is more crucial than buoyancy). Further, VIDEOPHY [5] focuses on semantic adherence and physical commonsense on video generation for diverse material types and their interactions. However, it does not provide insights into the physical law violations in the videos. We note the difference between VIDEOPHY-2 and existing work in Appendix Table 7.

To address these gaps, we propose VIDEOPHY-2, a challenging physical commonsense evaluation dataset for real-world actions. Specifically, we curate a list of **197 actions** across diverse physical activities (e.g., hula-hooping, playing tennis, gymnastics) and object interactions (e.g., bending an object until it breaks). Then, we generate 3940 detailed prompts from these seed actions using a large language model (LLM). Further, these prompts are used to synthesize videos with modern video generative models. Finally, we compile a list of **candidate physical rules** (and laws) that should be satisfied in the generated videos, using vision-language models in the loop. For example, in a video of *sportsperson playing tennis*, a physical rule would be that *a tennis ball should follow a parabolic trajectory under gravity*. For gold-standard judgments, we ask human annotators to

Table 1: **Data statistics.** We present the number of instances for diverse features (e.g., captions,) in the VideoPhy-2.

Feature	Number
Captions	3940
Unique actions	197
Generated videos	6800
Human annotations	102K
Avg. words in original caption	16
Avg. words in upsampled caption	138
Category	*# Actions*
Sports and Physical Activities	143
Object Interactions	54
Hard subset	60

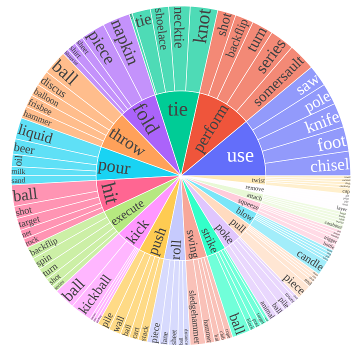


Figure 4: **Diversity of VideoPhy-2 prompts.** Top-20 frequently occurring verbs (inner) and their top-5 direct nouns (outer).

score each video based on overall semantic adherence and physical commonsense, and to mark its compliance with various physical rules. We present the entire pipeline in Figure 3.

In our experiments, we find that the best-performing model, Wan2.1-14B [55], achieves a joint performance score (high semantic adherence and physical commonsense) of only 32.6%. To further increase the dataset's difficulty, we create a **hard subset** which decreases the performance of Wan2.1-14B drops from 32.6% to 22%. Furthermore, our fine-grained analysis of human-annotated physical rule violations reveals that video models struggle the most with *conservation laws*, such as those governing mass and momentum. Overall, we demonstrate that VideoPhy-2 is a high-quality dataset that presents a formidable challenge for modern video models.

While human evaluation serves as the gold standard for real-world physical commonsense judgment, it is expensive and difficult to scale. To address this, we train an automatic evaluation model, **VideoPhy-2-AutoEval**, capable of performing a wide range of tasks—including scoring semantic adherence, physical commonsense, and classifying physical rule grounding in the generated video. In our experiments, we find that VideoPhy-2-AutoEval outperforms a capable multimodal foundation model, Gemini-2.0-Flash-Exp [15], with a relative correlation improvement of 81% and 236% on the semantic adherence and physical commonsense tasks, respectively, on the unseen prompts. Overall, our dataset represents a significant improvement over prior work and lays the foundation for assessing next-generation physical simulators on real-world tasks.

2 VideoPhy-2 Dataset

In this work, VideoPhy-2 aims to assess the ability of modern text-to-video generative models to adhere to the input prompt and generate physically realistic videos. We present the steps for data construction below:

Seed Actions (Stage 1): First, we curate a set of actions relevant to physical commonsense evaluation. Specifically, we compile a diverse list of over 600 actions from popular video datasets that capture a wide range of real-world activities, particularly those involving sports, physical activities, and object interactions. These datasets include Kinetics [13], UCF-101 [50], and SSv2 [19]. Next, we divide the student authors, with undergraduate or more degree in STEM, into two groups, each of which independently reviews the list and marks actions deemed relevant for physical commonsense evaluation. Our goal is to include actions that test various physical laws (e.g., gravity, elasticity, buoyancy, reflection, conservation of mass and momentum). Importantly, we filter out actions that do not elicit significant motion or are unlikely to be compelling for physical commonsense evaluation in videos (e.g., typing, applying cream, arguing, auctioning, chewing, playing instruments, petting a cat). Finally, we retain only the actions deemed relevant by both groups of annotators. After this filtering process, we obtain a list of 232 actions, which we further refine using Gemini-2.0-Flash-Exp to remove semantic duplicates, resulting in a final set of **197** actions. Among these, 54 actions focus on object interactions, while 143 pertain to physical and sports activities. We present the list of all the actions in Appendix Table 9.

LLM-Generated Prompts (Stage 2): In this stage, we query the Gemini-2.0-Flash-Exp LLM to independently generate 20 prompts for each action in our dataset. Specifically, the prompt generation follows several key principles: (a) focus on visible physical interactions between objects that can be clearly grounded in a generated video (e.g., an arrow hitting a target), (b) exclusion on non-visual details, such as mental states (e.g., emotions and intent), sensory details (e.g., smells and sounds), and abstract or poetic language that does not translate into a clear visual representation, (c) incorporating diverse characters and objects, and (d) depiction of multiple events within a prompt to increase the challenge for modern video generation models (e.g., we encourage the LLM to generate 'An archer draws the bowstring back to full tension, then releases the arrow, which flies straight and strikes a bullseye on a paper target' instead of a simpler prompt 'An archer releases the arrow'). Our prompt generation template is presented in Appendix E. In total, we curated 3940 prompts covering a wide range of actions. Since the modern video models can understand long video descriptions, we also generate dense captions from the original captions using the Mistral-NeMo-12B-Instruct prompt upsampler from [1]. In particular, these dense captions add more visual details to the original caption without changing its semantic meaning (e.g., main characters and actions). We present some of the generated captions and underlying actions in Appendix Table 11, and the upsampled captions in Appendix Table 12.

Candidate physical rules and laws (Stage 3): In this stage, we aim to generate candidate physical rules and associated laws that could be followed (or violated) in the generated video. Since video models often struggle to adhere to conditioning text prompts, we do not derive physical rules directly from them. Instead, we first generate videos using generative models conditioned on prompts from the VIDEOPHY-2 dataset. Then, we create captions for these videos using the strong video captioning capabilities of Gemini-2.0-Flash-Exp. This ensures that the physical rules are constructed based on details grounded in the video itself.[1] Subsequently, we ask Gemini-2.0-Flash-Exp to generate a set of three physical rules (and laws) that should be followed for a given video. Since a video may violate physical rules that are not covered in the pre-defined rules, we further ask the human annotators to write additional violated rules during physical commonsense evaluation. We present the rule generation prompt in Appendix Table 12.

Construction of the Hard Subset (Stage 4): While we collect diverse and lengthy captions to make the task more challenging, we further employ a model-based strategy to identify a subset of particularly difficult actions. Specifically, we generate videos using a strong open video model, CogVideoX-5B [58], conditioned on captions from the VIDEOPHY-2 dataset. From this, we select 60 actions (out of 197) for which the model fails to generate videos that accurately adhere to the prompts and follow physical commonsense (Appendix Table 10). On examination, we find that these actions focus on physics-rich interactions (e.g., momentum transfer in throwing discus or passing football), state changes (e.g., bending something until it breaks), balancing (e.g., tightrope walking), and complex motions (e.g., backflip, pole vault, and pizzatossing). In total, we designate 1200 prompts making the dataset more challenging. We present the list of hard actions in Appendix Table 10.[2]

Data Analysis: We present the dataset statistics in Table 1. Specifically, VIDEOPHY-2 contains 3940 captions, which is $5.72\times$ more than those in the VIDEOPHY dataset. Additionally, the average lengths of original and upsampled captions are 16 and 138 tokens, respectively—$1.88\times$ and $16.2\times$ longer than those in VIDEOPHY. Furthermore, VIDEOPHY-2 includes 102K human annotations across various video generative models and their semantic adherence, physical commonsense, and physical rule annotations. Finally, we show the distribution of the root verbs and direct nouns in the original captions of VIDEOPHY-2 in Figure 4, demonstrating the high diversity of the dataset. We also illustrate the diversity of multiple captions for a specific action in Appendix Figure 8. Overall, our analysis highlights that VIDEOPHY-2 significantly enhances data diversity and richness compared to its counterparts.

[1] We observe that prompting Gemini-2.0-Flash-Exp to generate physical rules directly from the video did not yield high-quality outputs. Therefore, we prefer a two-step process: first captioning the video, then generating the rules.

[2] We note that a similar model-based strategy is also adopted in recent works like Humanity's Last Exam [46] and ZeroBench [49] to collect hard instances for model evaluation.

3 Evaluation

3.1 Metric

In practice, we want the generated video to adhere to several constraints, including high video quality [37], temporal consistency [25], entity and background consistency [4]. While many of these metrics are intertwined, it is crucial to evaluate each one independently to gain a better understanding of the model's capabilities. In this regard, we focus on the extent to which the generated video adheres to the input text prompt and follows physical commonsense. Similar to [5, 36, 37], we prioritize rating-based feedback, as it quantifies the mode of failure or success for individual videos. In contrast, ranking-based feedback does not measure the magnitude of the difference between two videos but simply indicates which one is preferred. Unlike prior work [5], we collect dense rating feedback on a 5-point scale, allowing human annotators to express their judgments in a more fine-grained manner. Furthermore, we extend previous studies by evaluating whether generated videos adhere to diverse physical rules and their laws.

Semantic Adherence (SA): Here, we aim to assess whether the input text prompt is semantically grounded in the generated video. Specifically, it studies whether the entities, actions, and relationships described in the prompt are accurately depicted in the video (e.g., a person visibly jumping into the water). To measure semantic adherence, annotators rate each video on a 5-point scale, selecting from the following options: $\{SA \in$ *Very Unlikely (1), Unlikely (2), Neutral (3), Likely (4), Very Likely (5)*$\}$. In this case, *very unlikely* indicates that the video does not match the prompt at all, and *very likely* highlights the video fully adheres to the prompt with no inconsistencies.

Physical Commonsense (PC): Here, our goal is to assess whether the generated video follows the physical laws of the real-world intuitively (e.g., the football should start moving after impact in accordance with newton's first law). We note that the physical commonsense evaluation is independent of the underlying video generating text prompt. Since a video can follow (or violate) numerous laws, we are concerned with the holistic sense of the video's physical commonsense. In particular, the annotators rate each video on a 5-point scale, selecting from the following options: $\{PC \in$ *Very Unlikely (1), Unlikely (2), Neutral (3), Likely (4), Very Likely (5)*$\}$. Here, *very unlikely* that the video contains numerous violations of fundamental physical laws, and *very likely* indicates that the video demonstrates a strong understanding of physical commonsense with no violations.

Similar to [5], we compute **joint performance** as the main evaluation metric, which measures the fraction of videos that both adhere closely to the text prompt ($SA \geq 4$) and follow physical commonsense to a high degree ($PC \geq 4$). We do not report the posterior score ($PC >= 4 | SA >= 4$) since a bad model can game it.[3]

Physical Rules (PR): A key feature of the VIDEOPHY-2 dataset is the collection of candidate physical rules (and their associated laws) that humans evaluate as being followed or violated in the generated video (e.g., 'the ball should go down' is a physical rule associated with the law of gravity). These rules enable a fine-grained assessment of the video model's capabilities. Specifically, we determine whether a candidate physical rule is *violated (0)*, *followed (1)*, or *cannot be determined (2)* in the generated video.[4] Further, we ask human annotators to note more physical rule violations to ensure comprehensive coverage.

3.2 Human Evaluation

In practice, human evaluation serves as a gold standard for assessing the quality of generative foundation models [39, 59]. In particular, we collect judgments using the Amazon Mechanical Turk (AMT) platform from a group of 12 human annotators, which were selected after passing a qualification test. To promote high-quality annotations, we communicated with annotators remotely to provide detailed instructions and analyze annotation examples.[5] Since physical commonsense is in-

[3]For example, a model can adhere to the prompt for 1 out of 1000 prompts in the dataset. Now, assume that this video is also physically realistic. Then, the posterior performance of this model will be 100% that is quite misleading for the model builders.

[4]We include CBD category because LLM-generated physical rules may not always be visually grounded in the video.

[5]We pay a wage of $18 per hour to our annotators.

dependent of the generated video-prompt alignment, we evaluate semantic adherence and physical commonsense (including rule-based judgment) as separate tasks for human annotators. This differs from prior work in VIDEOPHY [5], which treats semantic adherence and physical commonsense assessment as a single task. It may introduce evaluation bias, as annotators have access to the prompt while conducting the physical commonsense evaluation, a scenario we explicitly avoid in this work.

We present the annotation UI for the semantic adherence task in Appendix Figure 18, where the input consists of a text prompt and the corresponding generated video. Note that human annotators were shown the original prompt (not the upsampled prompts) to ensure a fair comparison between video models, regardless of their ability to handle short or long prompts. In the following task, human annotators are asked to evaluate only the generated video and with regard to adherence to specific physical rules (followed/violated/cannot be determined), overall physical commonsense (rated on a scale of 1-5), and observable behaviors that violate physical reality.[6] The annotation interface for this task is shown in Appendix Figure 19.

3.3 Automatic Evaluation

While human judgments serve as the gold standard, automating the evaluation process is crucial for faster and more cost-effective model assessments. In this study, we evaluate several video-language foundation models (e.g., Gemini-2.0-Flash-Exp, VideoScore [20]) on two tasks: semantic adherence and physical commonsense scoring. Specifically, we prompt the models to score generated videos based on these two criteria and then normalize their predictions to a 5-point scale. We provide more details about score computation in Appendix J. Additionally, we introduce a classification task to determine whether a given physical rule is followed, violated, or indeterminate (CBD) in the generated video, leveraging video-language models such as VideoLLaVA [34]. In this task, we prompt the model to classify each video-rule pair into one of three categories: followed, violated, or CBD.

Our experiments reveal that existing video-language models struggle to achieve strong agreement with human annotators. This discrepancy primarily arises due to their limited understanding of physical commonsense and rules, as well as the complexity of the prompts. Hence, we supplement our benchmark with a video-language model VIDEOPHY-2-AUTOEVAL (7B parameters). Specifically, we aim to provide more accurate predictions for the generated videos along three axis – semantic adherence score (1-5), physical commonsense score (1-5), and physical rule classification (0-2). We follow a data-driven approach to distill human knowledge into a foundation model for these tasks. Specifically, we fine-tune a video-language model VideoCon-Physics [5] on 50K human annotations acquired for these tasks. We train a mult-task model to solve the three tasks using a shared backbone, to allow the inter-task knowledge transfer. We provide the templates and setup used for model finetuning in Appendix I and Appendix H, respectively.

4 Setup

Video generative models. In this work, we evaluate a diverse range of state-of-the-art text-to-video generative models. Specifically, we assess five open models and two closed models, including *CogVideoX-5B* [58], *VideoCrafter2* [14], *HunyuanVideo-13B* [31], *Cosmos-Diffusion-7B* [1], *Wan2.1-14B* [55], *OpenAI Sora* [10], and *Luma Ray2* [40].[7] We prompt these models with the upsampled captions, except for those that do not support long (dense) captions. Specifically, Hunyuan-13B and VideoCrafter2 are limited to 77 tokens due to their reliance on the CLIP [48] text encoder. Additionally, we generate short videos (less than 6s) as they are easier to evaluate and effectively highlight challenges on the VIDEOPHY-2. The model inference details are provided in Appendix L.

Dataset setup. Similar to [5], we take a data-driven approach and use human annotations across multiple tasks to train the automatic evaluator. We split the VIDEOPHY-2 dataset into a test set for benchmarking and a training set for training the VIDEOPHY-2-AUTOEVAL model. Specifically, the training and testing prompts consist of 3350 (197 actions × 17 captions per action) and 590 (197 actions × 3 captions per action) prompts, respectively.

[6] In our instructions to the annotators, we explicitly clarify that the overall physical commonsense judgments should extend beyond the predefined physical rules listed in the task.

[7] We exclude other closed models due to lack of API access (e.g., Veo2 [53], Kling [29]).

Table 2: **Human evaluation results on VIDEOPHY-2.** We present the joint performance that focuses on high semantic adherence and high physical commonsense in the generated videos. Hard, PA, OI refer to the hard, physical activities, and object interactions subsets of the data, respectively. We mark the best performing models in each column by blue and second best by yellow.

Model	Class	All	Hard	PA	OI
Wan2.1-14B [55]	Open	32.6	21.9	31.5	36.2
CogVideoX-5B [58]	Open	25.0	0.0	24.6	26.1
Cosmos-Diff-7B [1]	Open	24.1	10.9	22.6	27.4
Hunyuan-13B [31]	Open	17.2	6.2	17.6	15.9
VideoCrafter-2 [14]	Open	10.5	2.9	10.1	13.1
Ray2 [40]	Closed	20.3	8.3	21.0	18.5
Sora [10]	Closed	23.3	5.3	22.2	26.7

Benchmarking. For every tested model, we generate one video per each test prompt, that is, 590 videos per model. For Sora, however, we generate a subset of 60 videos (randomly selected from 590), manually, using Sora playground[8] due to the lack of an official API, and 394 videos (2 prompts per action) for Ray2 due to the limited API budget. After generating the videos, we ask three annotators to evaluate them based on semantic adherence, overall physical commonsense, and violations of various physical rules. Annotators can also suggest additional physical rules that may be missing from our list. We observe that annotator agreement ranges from 75% to 80% for these tasks, which is reasonable given the task subjectivity and comparable to the agreement scores in prior work [5]. For every generated video, we compute the SA and PC scores (1-5) by averaging the three annotators scores and rounding to the nearest integer. Following this, the joint score is computed to assess the quality of the generated video. We use the majority voting for determining whether the listed physical rule (and law) is followed, violated, or cannot be grounded in the generated video. Additional human-written violations are converted to a statement of a physical rule (and law) using Gemini-2.0-Flash-Exp. With CogVideoX-5B as a strong reference model, we choose a *hard* subset of 60 actions for which it achieved a zero joint performance. In our experiments, we observe that this hard subset leads to big drop in performances in comparison to the entire data across diverse video models. In total, we collect 10.2K, 10.2K, and 30.6K semantic adherence, physical commonsense, and physical rule annotations, which cost us $2600 USD.

Training set for VIDEOPHY-2-AUTOEVAL. In this case, we sample 1 video per caption from one of the three capable video models (HunyuanVideo-13B, Cosmos-Diffusion-7B, and CogVideoX-5B) from the training set, of size 3350. Subsequently, we perform human annotations in the same way as the benchmarking process i.e., aggregating semantic adherence, physical commonsense and rule judgments across the three annotators. In total, we collect $\sim 50K$ human annotations across the three tasks, and spend $3515 USD on collecting the training data. Post-training, we compare the performance of VIDEOPHY-2-AUTOEVAL against several baselines on the semantic adherence and physical commonsense judgments using Pearson's correlation between the ground-truth and predicted scores. Further, we compare the joint score prediction accuracy and F1 score between our auto-rater and selected baselines. In addition, we compare the physical rule classification accuracy between the VIDEOPHY-2-AUTOEVAL and baselines.

Table 3: **Correlation analysis between semantic adherence and physical commonsense with other video metrics.**

	Aesthetics	Motion	SA
SA	0.1	0.02	1
PC	0.09	0.002	0.14

5 Experiments

Here, we present the benchmarking results and the fine-grained analysis (§5.1). Then, we note the usefulness of our auto-rater against modern video-language models (§5.2).

[8] https://openai.com/sora/

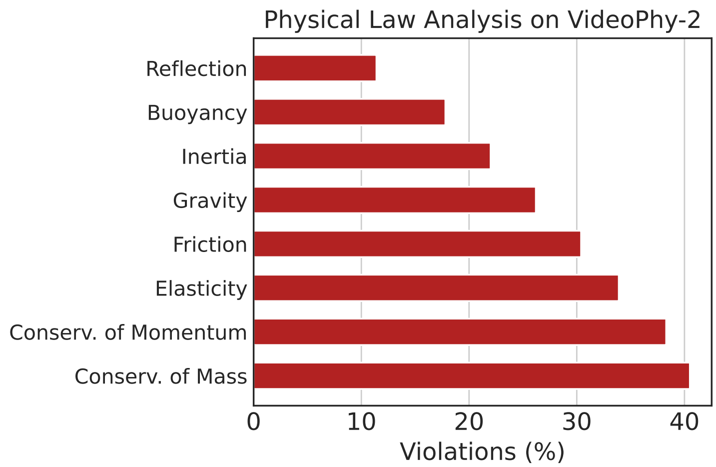


Figure 5: **Physical laws violation analysis.** We present the violation scores for diverse physical laws based on human annotations collected from various video generative models on VIDEOPHY-2.

(a) **v. Ray2**, Two jetskis race side-by-side, their wakes colliding in a visible spray of water.

(b) **v. Hunyuan**, A sledgehammer is swung at a wooden fence post, causing it to break and fall to the ground.

(c) **v. Cosmos-1.0**, A javelin thrown from a high vantage point lands in a rocky outcrop below, with visible dust.

Figure 6: **Comparison of Wan2.1 with other models**. The top row shows videos generated by Wan2.1: (a) For Ray2, the jetski on the left lags behind the other jetski and then starts moving backward. (b) For Hunyuan-13B, the sledgehammer deforms after the swing, and a broken wooden board appears out of nowhere. (c) For Cosmos-7B, the javelin expels sand before it even hits the ground.

5.1 Main Results

Performance on the dataset. We compare the joint performance of various open and closed text-to-video generative models on the VIDEOPHY-2 dataset in Table 2. Specifically, we present their performance on the entire dataset, the hard split, and subsets focused on physical activities/sports (PA) and object interactions (OI). Even the best-performing model, Wan2.1-14B, achieves only 32.6% and 21.9% on the full and hard splits of our dataset, respectively. Its relatively strong performance compared to other models can be attributed to the diversity of its multimodal training data, along with robust motion filtering that preserves high-quality videos across a wide range of actions.

Furthermore, we observe that closed models, such as Ray2, perform worse than open models like Wan2.1-14B and CogVideoX-5B. This suggests that closed models are not necessarily superior to open models in capturing physical commonsense. Notably, Cosmos-Diffusion-7B achieves the second-best score on the hard split, even outperforming the much larger HunyuanVideo-13B model. This may be due to the high representation of human actions in its training data, along with synthetically rendered simulations [1].

Additionally, we find that performance on physical activities (sports) is generally lower than on object interactions across different video models. This suggests that future data curation efforts should focus on collecting high-quality sports activity videos (e.g., tennis, discus throw, baseball, cricket) to improve performance on the VIDEOPHY-2 dataset. Finally, we present the correlation between SA and PC judgments and other video metrics, including aesthetics (measured using the LAION classifier [32]) and motion quality (measured using optical flow from RAFT [51]), in Table 3. Our results reveal that physical commonsense is not well-correlated with any of these video metrics. This

(a) Someone inflates a water balloon until it is taut and about to burst, the rubber stretching visibly.

(b) A small rock tumbles down a steep, rocky hillside, displacing soil and small stones.

Figure 7: **Illustration of Wan2.1's bad physical commonsense**. Even the best-performing model, Wan2.1, may struggle to correctly capture physical laws, leading to the generation of unnatural videos. Examples of such artifacts include: (a) A hot air balloon, which should be full of air, instead shrinking and expelling water. (b) A rock rolling and accelerating uphill instead of downhill.

Table 4: **Auto-rater evaluation results.** We present the pearson's correlation ($\times$100) between the predicted scores and ground-truth scores (1-5) on the unseen prompts and unseen video models.

	Unseen prompts			**Unseen video models**		
	Avg.	**SA**	**PC**	**Avg.**	**SA**	**PC**
VideoCon-Physics [5]	28.5	32.0	25.0	26.5	27.0	26.0
VideoCon [3]	12.5	23.0	2.0	8.9	17.0	0.8
VideoLlava [34]	16.0	30.0	2.0	19.0	33.0	5.0
VideoScore [20]	13.5	17.0	10.0	9.0	5.0	13.0
Gemini-2.0-Flash-Exp	18.5	26.0	11.0	21.0	31.0	11.0
VIDEOPHY-2-AUTOEVAL	42.0	47.0	37.0	41.0	45.0	37.0
Rel. to Best (%)	+47.4	+46.9	+48.0	+49.0	+36.4	+61.5
Rel. to Gemini (%)	+127.0	+80.8	+236.4	+107.1	+45.2	+281.8

indicates that a model cannot achieve high performance on our dataset simply by optimizing for aesthetics and motion quality; rather, it requires dedicated efforts to incorporate physical commonsense into video generation. Overall, our findings suggest that VIDEOPHY-2 presents a significant challenge for modern video models, with substantial room for improvement in future iterations.

Fine-grained Analysis. In our human annotations, we create a list of physical rules (and associated laws) that are violated in each video of the VIDEOPHY-2 dataset. We then analyze the fraction of instances in which a physical law is violated to gain fine-grained insights into model behavior. For example, if 100 physical rules are associated with the law of gravity and 25 of them are violated, the violation score would be 25%. We present the results of physical law violations in Figure 5. We observe that the conservation of momentum (linear or angular) and the conservation of mass are among the most frequently violated physical laws, with violation scores of 40%, in the videos from the VIDEOPHY-2 dataset. Conversely, we find that reflection and buoyancy are relatively mastered with violation scores less than 20%.

Qualitative Analysis. We present qualitative analysis to provide visual insights into the model's mode of failures. Specifically, we cover model-specific poor physical commonsense instances along the caption and human-judged physical violations in Appendix N. For example, we show that the Sora-generated video violates the physical rule 'The frisbee must contant the hand before any upward movement occurs' (Appendix Figure 2). We also provide several qualitative examples across diverse physical law violations across different models in Appendix O. For example, we highlight that the 'golf ball does not move after being struck by the golf club' for Ray2 (Figure 26). Furthermore, we present qualitative examples in Figure 6 to compare the best-performing model, Wan2.1-14B, with other video models. Notably, we observe violations of physical commonsense, such as jetskis moving unnaturally in reverse and the deformation of a solid sledgehammer, defying the principles of elasticity. However, even Wan suffers from the lack of physical commonsense, as shown in Figure 7. In this case, we highlight that a rock starts rolling and accelerating uphill, defying the physical law of gravity.

5.2 VIDEOPHY-2-AUTOEVAL

To enable scalable judgments, we supplement the dataset with an automatic evaluator VIDEOPHY-2-AUTOEVAL. We assess the auto-rater performance for two settings: (a) unseen prompts: where the auto-rater is assessed on the same video models as used in training but on the videos generated

Table 5: **Auto-rater evaluation on joint score judgments.** We present the joint accuracy and F1 score between the predicted scores and ground-truth scores (0-1) for our VIDEOPHY-2-AUTOEVAL and VideoCon-Physics.

Method	Unseen prompts			Unseen video models		
	Avg.	Acc.	F1	Avg.	Acc.	F1
VideoCon-Physics [5]	39.1	75.6	2.6	39.6	75	4.2
VIDEOPHY-2-AUTOEVAL	65.1	79.1	51.1	62.8	76.3	49.3
Rel. to VideoCon-Physics (%)	+66.4			+49.1		

Table 6: **Auto-rater evaluation on physical rule classification.** We present the accuracy results for VIDEOPHY-2-AUTOEVAL and other video-language models on the rule classification tasks.

	Unseen prompts	Unseen video models
Random	34.5	31.2
VideoLlava [34]	38.1	38.7
Gemini-2.0-Flash-Exp	59.2	57.1
VIDEOPHY-2-AUTOEVAL	78.7	72.9
Rel. to Best (%)	+32.9	+27.7

for unseen (testing) captions, (b) unseen video models: where the auto-rater is assessed on the videos generated from the unseen video models for unseen (testing) captions.

We compare the correlation performance of VIDEOPHY-2-AUTOEVAL against several baselines in Table 4. In particular, VIDEOPHY-2-AUTOEVAL achieves relative gains of 47.4% and 49% on unseen prompts and unseen video models, respectively, compared to the best-performing baselines. Further, our auto-rater outperforms the state-of-the-art multimodal model, Gemini-2.0-Flash-Exp, with relative gains of 81% in semantic adherence and 236% in physical commonsense judgments. Further, we evaluate the accuracy and F1 performance of VIDEOPHY-2-AUTOEVAL against VideoCon-Physics for joint score judgments in Table 5. Our results show that VIDEOPHY-2-AUTOEVAL maintains a strong balance between joint accuracy and F1 scores. These findings highlight the need for further improvements in video-language models to enhance their physical commonsense understanding. Finally, we assess the physical rule classification accuracy of VIDEOPHY-2-AUTOEVAL against baselines in Table 6. Our model achieves relative gains of 32.9% on unseen prompts and 27.7% on unseen video models compared to Gemini-2.0-Flash-Exp. This demonstrates that our unified auto-rater can reliably handle a variety of tasks, providing future model developers with a robust tool for testing on the VIDEOPHY-2 dataset.

6 Conclusion

We introduce VIDEOPHY-2, a benchmark for evaluating physical commonsense in videos generated by modern models. We reveal a large gap in their ability to align with prompts and generate videos that follow physical commonsense. Further, we provide physical law violations and an auto-rater for scalable evaluation. Overall, this dataset advances our understanding of the current state of the video generative models as general-purpose world simulators.

7 Acknowledgements

Hritik Bansal is supported in part by AFOSR MURI grant FA9550-22-1-0380. We also thank Ashima Suvarna and Zongyu Lin for their helpful feedback on the draft.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3, 5, 7, 8, 9, 17

[2] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning, 2019. 17

[3] Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13927–13937, 2024. 10, 19

[4] Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szpektor, Aditya Grover, and Kai-Wei Chang. Talc: Time-aligned captions for multi-scene text-to-video generation. *arXiv preprint arXiv:2405.04682*, 2024. 6

[5] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 3, 6, 7, 8, 10, 11, 16, 17, 18, 19

[6] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. 17

[7] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. 17

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 16

[9] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023. 16

[10] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 3, 7, 8, 16

[11] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. 16

[12] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 3, 17

[13] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 4

[14] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 7, 8

[15] Google DeepMind. Introducing Gemini 2.0: our new AI model for the agentic era — blog.google. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`, 2025. 4

[16] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 17

[17] Jiafei Duan, Arijit Dasgupta, Jason Fischer, and Cheston Tan. A survey on machine learning approaches for modelling intuitive physics. *arXiv preprint arXiv:2202.06481*, 2022. 3, 17

[18] GoogleDeepMind Genie2. Genie 2: A large-scale foundation world model — deepmind.google. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/, 2024. 3

[19] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 4

[20] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 7, 10, 17, 19

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. 16

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 16

[23] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 16

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 18

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6, 16

[26] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators, 2023. 16

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 18

[28] Gergely Klár, Theodore Gast, Andre Pradhana, Chuyuan Fu, Craig Schroeder, Chenfanfu Jiang, and Joseph Teran. Drucker-prager elastoplasticity for sand animation. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 3

[29] KlingAI. KLING AI — klingai.com. https://www.klingai.com/, 2024. 7

[30] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. 16

[31] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 7, 8

[32] Laion. GitHub - LAION-AI/aesthetic-predictor: A linear estimator on top of clip to predict the aesthetic quality of pictures — github.com. https://github.com/LAION-AI/aesthetic-predictor. 9

[33] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024. 3

[34] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 7, 10, 11, 19

[35] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. 17

[36] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024. 3, 6

[37] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 6, 16

[38] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 16

[39] lmarena. Chatbot Arena Leaderboard - a Hugging Face Space by lmarena-ai — huggingface.co. https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard. 6

[40] LumaAI. Luma Ray2 — lumalabs.ai. https://lumalabs.ai/ray, 2025. 7, 8

[41] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation, 2025. 16

[42] Michael McCloskey, Allyson Washburn, and Linda Felch. Intuitive physics: the straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4):636, 1983. 17

[43] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 16

[44] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025. 3, 16

[45] James F O'brien, Adam W Bargteil, and Jessica K Hodgins. Graphical modeling and animation of ductile fracture. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 291–294, 2002. 3

[46] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. 5

[47] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024. 3, 17

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 19

[49] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025. 5

[50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[51] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 9

[52] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges, 2019. 17

[53] GoogleDeepMind Veo2. Veo 2 — deepmind.google. `https://deepmind.google/technologies/veo/veo-2/`, 2024. 3, 7

[54] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description, 2022. 16

[55] Wan. Wan-AI/Wan2.1-T2V-14B · Hugging Face — huggingface.co. `https://huggingface.co/Wan-AI/Wan2.1-T2V-14B`, 2025. 4, 7, 8

[56] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. 16

[57] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023. 16

[58] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 5, 7, 8

[59] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36:1601–1619, 2023. 6

Table 7: **Comparison between VIDEOPHY-2 and several prior work.** We highlight the salient features of the VIDEOPHY-2 and show that its unique contributions. For instance, it is one of the largest datasets for physical commonsense evaluation, along with violated physical rule (and law) annotations. Further, we will release all the data, and videos openly for public-use.

Feature	VBench[25]	PhyGenBench [43]	PhysicsIQ [44]	EvalCrafter [37]	VIDEOPHY [5]	VIDEOPHY-2 (Ours)
Num of captions	1746	160	396	700	688	3940
Gold human evaluation	✓	✓	✗	✓	✓	✓
Physical commonsense eval.	✗	✓	✓	✗	✓	✓
Physical rules and laws annotations	✗	✓	✗	✗	✗	✓
Real-world action-centric	✗	✗	✗	✗	✗	✓
Long (dense) captions	✗	✓	✓	✗	✗	✓
Hard subset	✗	✗	✗	✗	✗	✓
Automatic evaluator	✓	✓	✓	✓	✓	✓
Release videos and annotations	✓	✗	✗	✓	✓	✓
Human feedback type	Pairwise	Rating	-	Rating (1-5)	Rating (0-1)	Rating (1-5)

A Related Work

The rapid advancement of video generation models necessitates robust benchmarks and evaluation methodologies, particularly for assessing *physical commonsense* understanding. This is a crucial step towards realizing the vision of video generation models as 'world simulators' [10]. Our work, VIDEOPHY-2, extends prior work by addressing key limitations in existing benchmarks and evaluation methods. We structure our discussion of related work around the following key areas:

A.1 Video Generation Models

Recent progress in text-to-video (T2V) generation has been driven primarily by two architectural paradigms: diffusion models [21, 22, 9, 57, 56, 26] and autoregressive models [41, 30, 23, 54]. Diffusion models, such as Stable Video Diffusion (SVD) [8] and Sora [38], involve a multi-stage training process and operate in a latent space. Autoregressive models, including VideoPoet [30] and CogVideo [23], predict future frames based on past frames. While these models demonstrate impressive visual fidelity, their ability to capture underlying physical principles remains an open question. This highlights the need for rigorous benchmarks like VIDEOPHY-2. Additionally, recent works, such as Genie [11], explore interactive video generation, further expanding the potential applications of these models beyond static content creation.

A.2 General Video Generation Benchmarks

Several benchmarks focus on evaluating general aspects of video quality but do not specifically assess physical reasoning. **VBench** [25] introduces a hierarchical evaluation approach, covering motion smoothness, background consistency, and overall visual fidelity. **EvalCrafter** [37] proposes 17 objective metrics focused on different aspects of video quality. While these benchmarks contribute to understanding model performance, they do not isolate or systematically evaluate physical commonsense in generated videos.

A.3 Benchmarks for Physical Understanding in Video Generation

Several works have introduced benchmarks to assess the physical plausibility of generated videos. **VIDEOPHY** [5] was an early effort in this direction, evaluating semantic adherence and physical plausibility across 688 prompts. VIDEOPHY-2 extends this work by introducing a more extensive dataset of 3940 prompts, shifting the focus from material interactions to *real-world actions*, and providing a *5-point Likert scale* for human evaluation. Furthermore, VIDEOPHY-2 explicitly annotates physical rules in each video, allowing for a more fine-grained understanding of model behavior.

Physics-IQ [44] conditions models on initial frames of real videos and measures the similarity between predicted and ground-truth video continuations. While this approach offers valuable insights, it is primarily designed for video prediction rather than open-ended text-to-video generation. Additionally, extending it to longer or more complex multi-event scenarios presents challenges.

PhyGenBench [43] proposes a benchmark of 160 prompts and an automated evaluation framework, PhyGenEval. This work introduces structured evaluations of physical reasoning; however, its relatively small scale makes generalization difficult. Additionally, it adopts a strict one-to-one mapping between prompts and physical laws, whereas real-world physics often involves multiple interacting principles. Its evaluation relies on sequential vision-language model queries, which can introduce inconsistencies and increase computational complexity. VIDEOPHY-2 builds on these efforts by expanding dataset size, allowing for prompts that reflect multiple physical laws, and providing a more interpretable evaluation framework.

WorldSimBench [47] assesses video models' ability to act as *world simulators* by aligning their outputs with numerical solvers. While valuable, this approach faces challenges in bridging the *sim-to-real gap*, as physical simulations may not fully capture the complexity and variability of real-world interactions. VIDEOPHY-2 addresses this by exclusively using real-world videos to ensure that evaluations remain closely aligned with practical physical dynamics.

A.4 Automatic Evaluation Methods

Traditional video quality metrics, such as Fréchet Video Distance (FVD) [52], were not designed to assess physical plausibility. More recent approaches incorporate vision-language models (VLMs) for automatic evaluation [20, 5, 35]. VIDEOPHY introduced a fine-tuned VLM-based evaluator, but existing VLMs still face challenges in reliably assessing physical commonsense. VIDEOPHY-2 proposes VIDEOPHY-2-AutoEval, an enhanced automatic evaluator trained on a larger and more diverse dataset with explicit physical rule annotations. This leads to improved alignment with human assessments while maintaining interpretability.

A.5 Physical Reasoning in AI

The study of physical commonsense in AI builds upon insights from both artificial intelligence and cognitive science. Research on intuitive physics has explored how humans reason about object interactions and causal physical events [42, 6, 17]. In AI, approaches such as PIQA [7] assess physical reasoning in language, while others focus on synthetic environments for learning physical interactions [2, 16, 12, 1]. VIDEOPHY-2 contributes to this area by providing a large-scale, real-world benchmark that bridges the gap between theoretical physical reasoning and practical video generation.

In summary, VIDEOPHY-2 advances the evaluation of physical commonsense in video generation by introducing a large-scale, action-centric benchmark, a fine-grained evaluation framework with explicit rule annotations, and a robust, human-aligned automatic evaluator. These contributions address key limitations of prior work and facilitate a more comprehensive assessment of video models' ability to simulate the physical world.

B List of Actions

Our benchmark includes a broad set of actions spanning simple motions (e.g., *walking through snow, dribbling basketball*), object interactions (e.g., *peeling fruit, using a paint roller*), and complex physical activities (e.g., *pole vault, tightrope walking*). These actions ensure models are tested across diverse movement dynamics, real-world interactions, and varying levels of physical complexity.

Table 9 presents the full set of 197 actions.

C List of Hard Actions

Certain actions pose significant challenges for generative models due to rapid motion (e.g., *throwing discus, gymnastics tumbling*), intricate interactions (e.g., *popping balloons, pouring until overflow*), and structural deformations (e.g., *ripping paper, bending until breaking*).Actions were deemed hard using CogVideoX-5B as a baseline.

Table 10 presents a subset of 60 hard actions, ensuring our benchmark tests model capabilities beyond simple motions, including stability, object manipulation, and real-world physics.

D Diversity of prompts for diverse actions

In this work, our objective was to curate a set of actions from which we could generate a wide range of diverse prompts that, while focusing on motion, still encompass various real-world settings. By ensuring diversity in prompt selection within each action, our dataset becomes more robust and capable of testing generative models across a broad spectrum of physical interactions.

Figure 8 illustrates a subset of actions used in our study, along with the verb-noun pairs appearing in their respective generated prompts. The diversity of nouns and verbs in these prompts ensures that our dataset covers a wide range of contextual variations. For instance, the action *Rowing* can be associated with nouns such as *boat*, *paddle*, or *water*, leading to prompts that span different environments and interactions. Similarly, *Knitting* may involve *yarn*, *needles*, or *fabric*, allowing for a richer set of generated scenarios that helps to better evaluate video-generation models.

E LLM-generated captions prompt

In this section, we discuss the prompt we gave Gemini-2.0-Flash-Exp to generate a list of diverse prompts for each action, displayed in Table 11.

F Video-specific physical rule generation prompt

In this section we discuss the prompt we gave Gemini-2.0-Flash-Exp to generate a list of 3 rules for each prompt, displayed in Table 12.

G Upsampled Caption Examples

To enhance video generation quality, we upsample captions by upsampling them to make them more specific with finer details. This process helps models generate better videos by providing additional cues about motion, environment, and object interactions.

Table 12 presents examples where simple captions (e.g., *"A person uses nunchucks to break a stack of wooden blocks"*) are expanded into more vivid descriptions (e.g., including details about lighting, camera angles, and material properties). These enriched captions guide models toward generating more coherent and contextually accurate videos.

H Training Details for VIDEOPHY-2-AUTOEVAL

We finetune VideoCon-Physics [5] (7B) that acts a strong base model for semantic adherence and physical commonsense evaluation. Specifically, we use low rank adaptation [24] applied to all the transformer blocks including QKVO, gate, up, and down weight matrices. In our experiments, we set $r, \alpha = 32$ and dropout=0.05. We finetune the base model for 3 epochs and pick the best checkpoint using the performance on the validation set. Further, we use Adam optimizer [27] with a linear warmup steps of 50 steps followed by linear decay. We perform a hyperparameter search over several peak learning rates $\{5e-5, 1e-4, 5e-4, 1e-3\}$ and found $5e-4$ worked the best. We use 4 Nvidia A6000 GPUs with a global batch size of 64.

I Multimodal Prompts for Automatic Evaluation

We prompt our model to generate a text response conditioned on the multimodal template $\mathcal{T}_t(x)$ for semantic adherence, physical commonsense, and rule tasks. Formally,

$$\mathcal{T}_t(x) = \begin{cases} \mathcal{T}_{SA}(V, C), & t = SA \\ \mathcal{T}_{PC}(V), & t = PC \\ \mathcal{T}_{R}(V, R), & t = RS \end{cases} \quad (1)$$

where t is either semantic adherence to the caption, physical commonsense, or the rule score, C is the conditioning caption, V is the generated video for the caption C, and R is the rule candidate. We provide the multimodal templates ($\mathcal{T}_{SA}(V, C)$, $\mathcal{T}_{PC}(V)$, and $\mathcal{T}_{R}(V, R)$). We compute the score from the model using simple autoregressive generation.

We present the prompts used for the this multimodal evaluation for semantic adherence evaluation in Figure 15, physical commonsense alignment in Figure 16, and rule scoring in Figure 17.

J Baselines Judgments

J.1 VideoPhysics and VideoCon

For our baseline comparisons, we obtained raw scores from VideoPhysics[5] and VideoCon[3] frameworks, which originally produced scores in the range of 0-1. To maintain consistency with our 1-5 rating scale, we normalized these scores using linear scaling. Specifically, we applied the transformation $score_{normalized} = score_{raw} \times 4 + 1$, followed by rounding to the nearest integer. This transformation maps the minimum possible score (0) to 1 and the maximum possible score (1) to 5, preserving the relative performance differences between models while making the scores directly comparable to our human evaluation ratings.

J.2 VideoScore Regression

For our implementation of VideoScore [20], we selected specific component metrics that align with our evaluation objectives. For SA, we utilized the Text-to-Video Alignment score, as it effectively measures the correspondence between video content and textual descriptions. For PC, we employed the Factual Consistency Score, which assesses the physical plausibility of events depicted in the videos.

J.3 VideoLlava

For VideoLlava [34], we used the same prompts shown in Appendix I. The model produced scores directly on our 1-5 scale following the evaluation criteria provided in the prompts.

J.4 Gemini

For Gemini-2.0-Flash-Exp, we leveraged its larger context window to provide more detailed and structured evaluation prompts, allowing us to specify evaluation criteria with greater granularity, and adding few-shot examples. This enabled more comprehensive assessments for semantic adherence (Figure 13) and physical commonsense (Figure 14).

K Human Annotation Interface

Figure 18 and Figure 19 showcase our human annotation interface. The interface is designed to facilitate both semantic adherence and physical commonsense evaluation. We combine rule scoring and physical commonsense assessment into a single task, allowing annotators to provide additional rules they believe are violated.

L Model Inference Details

Table 8 summarizes the inference settings for the video generation models used in our study. The table highlights key parameters such as resolution, frame rate, guidance scale, sampling steps, and precision.

In our experiments, all models except those with a token limit of 77 (due to CLIP [48] embeddings) used upsampled captions. Models like Wan2.1-14B and CogVideoX-5B particularly benefit from these richer descriptions, enhancing the quality of their generated videos.

We evaluated the models using original and upsampled captions. Different models use different schedulers, and use different default precision levels (bf16, fp16, fp32).

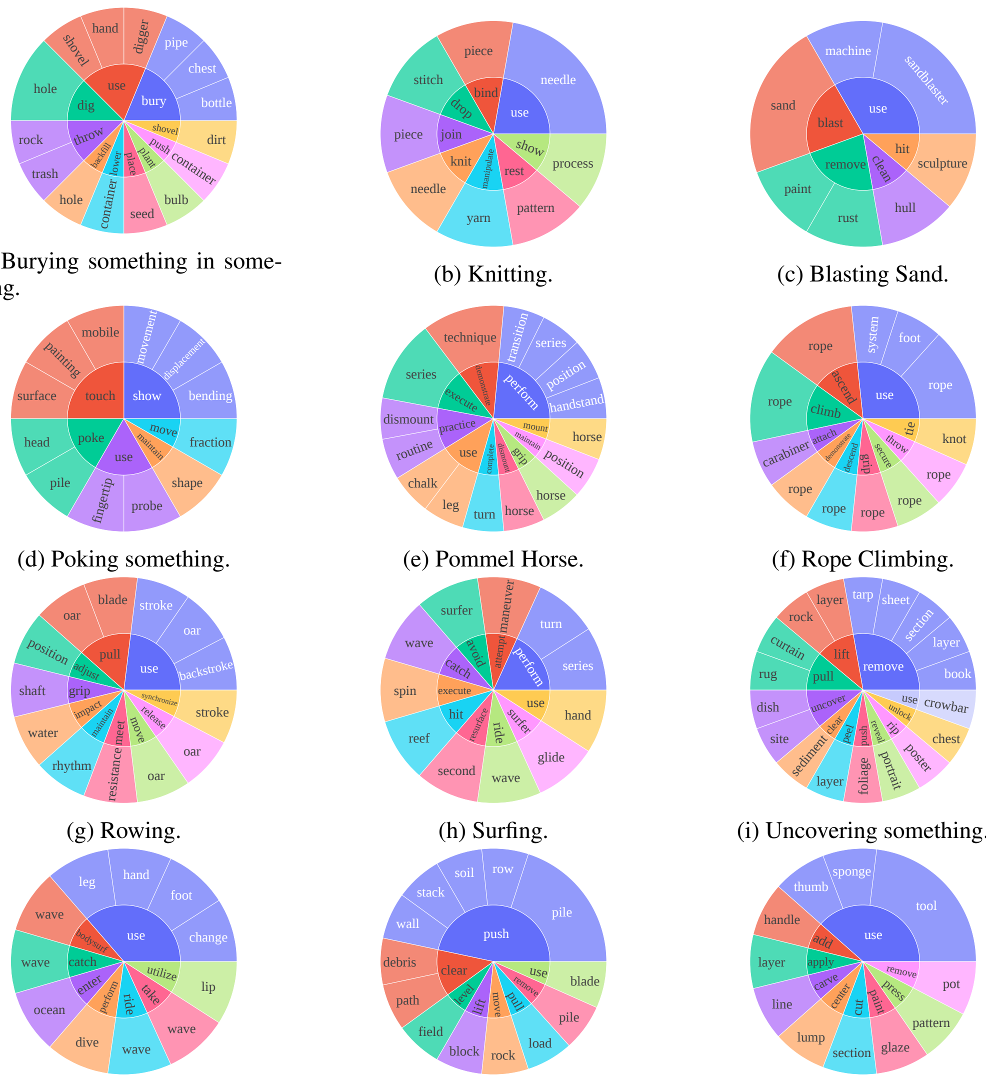
For closed models, such as Ray2 and Sora, we used fewer captions, which influences their evaluation outcomes.

M Distribution of semantic adherence and physical commonsense scores

We present the distribution of semantic adherence (SA) and physical commonsense (PC) scores in Figure 9 and 10.

Table 8: **Inference settings for different video generation models.**

Models	Caption Type	FPS/Video Length (s)	Resolution	Frames	Scale Guidance	Steps Sampling	Scheduler Noise	Precision
Wan2.1-14B	Upsampled	16 (4s)	832×480	61	5	50	FlowUniPCMultistepScheduler	bf16
CogVideoX-5B	Upsampled	8 (6s)	480×720	49	6	50	CogVideoXDPMScheduler	bf16
Cosmos-Diffusion-7B	Upsampled	24 (5s)	576×576	120	7	60	-	bf16
HunyuanVideo-13B	Original	15 (4s)	320×512	61	6	50	FlowMatchEulerDiscreteScheduler	fp16
VideoCrafter2-1.4B	Original	10 (3s)	320×512	32	12	50	DDIM	fp32



(a) Burying something in something.

(b) Knitting.

(c) Blasting Sand.

(d) Poking something.

(e) Pommel Horse.

(f) Rope Climbing.

(g) Rowing.

(h) Surfing.

(i) Uncovering something.

(j) Body surfing.

(k) Bulldozing.

(l) Clay pottery making.

Figure 8: A subset of actions, and the verb-noun pairs in their respective generated prompts.

N Poor physical commonsense qualitative examples by model

We present more examples from each generative model where one or more physical laws are violated in Figure 20 - Figure 25.

O Poor physical commonsense qualitative examples by law

We present a few qualitative examples highlighting instances where specific physical laws are violated in Figure 26 - Figure 33.

mopping floor, blowing out candles, throwing water balloon, passing american football, billiards, lifting a surface with something on it until it starts sliding down, using a sledge hammer, swing, hitting baseball, hammerthrow, playing tennis, longjump, sewing, roller skating, wading through water, riding mechanical bull, pole vault, blowdryhair, tying necktie, paragliding, something falling like a rock, playing ice hockey, sailing, gymnastics tumbling, pushing something so that it slightly moves, folding clothes, poking something so lightly that it doesn't or almost doesn't move, javelinthrow, surfing, snatch weight lifting, chiseling stone, spinning something that quickly stops spinning, poking a stack of something without the stack collapsing, carving ice, throwing something, twisting (wringing) something wet until water comes out, putting something onto something else that cannot support it so it falls down, wiping something off of something, playing field hockey, juggling balls, wading through mud, shooting basketball, welding, hoverboarding, javelin throw, catching or throwing softball, hammering, knitting, putting something that can't roll onto a slanted surface so it slides down, playing polo, pouring something onto something, pulling two ends of something so that it gets stretched, using circular saw, flint knapping, backflip (human), long jump, uncovering something, pouring something into something until it overflows, dribbling basketball, poking a hole into some substance, bulldozing, peeling fruit, parallelbars, playing darts, spinning something so it continues spinning, luge, pushing something so it spins, curling (sport), riding unicycle, throwing discus, folding paper, ripping paper, trapezing, playing pinball, burying something in something, throwing axe, wrapping present, yarn spinning, tying shoe laces, flying kite, tightrope walking, using a paint roller, using a wrench, sharpening pencil, pizzatossing, catching or throwing baseball, catching or throwing frisbee, playing kickball, golf, nunchucks, pouring something out of something, opening bottle (not wine), unfolding something, stuffing something into something, canoeing or kayaking, rolling something on a flat surface, playing ping pong, punching bag, picking fruit, poking something so that it spins around, balancebeam, parasailing, jumprope, bungee jumping, drop kicking, hammer throw, using segway, biking through snow, swimming, making snowman, rowing, attaching something onto something, something colliding with something and both come to a halt, blasting sand, throwdiscus, tying knot (not on a tie), digging something out of something, trimming shrubs, inflating balloons, bouncing on trampoline, spinning poi, shot put, pushing something onto something, something falling like a feather or paper, letting something roll down a slanted surface, cutting, chopping wood, breaststroke, hurdling, ice climbing, popping balloons, throwing knife, bending something so that it deforms, playing cricket, shaping bread dough, bobsledding, smashing, blowing bubble gum, something colliding with something and both are being deflected, breaking boards, somersaulting, skateboarding, squeezing something, ropeclimbing, bending something until it breaks, yoyo, bouncing on bouncy castle, hula hooping, letting something roll along a flat surface, pushing something off of something, lifting a surface with something on it but not enough for it to slide down, throwing snowballs, shoveling snow, rolling pastry, tossing coin, threading needle, skijet, clay pottery making, pommelhorse, playing squash or racquetball, pushing something so that it falls off the table, bodysurfing, twisting something, poking a hole into something soft, tearing something into two pieces, walking through snow, putting something that cannot actually stand upright upright on the table so it falls on its side, poking something so that it falls over, poking a stack of something so the stack collapses, pushing something so that it almost falls off but doesn't, kicking soccer ball, tying bow tie, wood burning (art), putting something that can't roll onto a slanted surface so it stays where it is, extinguishing fire, ice skating, playing badminton, archery, folding napkins, soccerjuggling, blowing leaves, bowling, mountain climber (exercise), jetskiing, pouring something into something, pouring beer, pulling two ends of something so that it separates into two pieces, volleyballspiking, poking something so it slightly moves, smoking, tearing something just a little bit, rock climbing, letting something roll up a slanted surface so it rolls back down, blowdrying hair, folding something, riding scooter, playing volleyball

Table 9: Actions List: A comprehensive collection of all physical actions that were used test video models' ability to produce videos that align with physical commonsense

blowing out candles, playing squash or racquetball, passing american football, billiards, backflip (human), tearing something into two pieces, pouring something into something until it overflows, lifting a surface with something on it until it starts sliding down, using a sledge hammer, swing, hitting baseball, peeling fruit, parallelbars, putting something that cannot actually stand upright upright on the table so it falls on its side, playing darts, poking something so that it falls over, chopping wood, throwing discus, poking a stack of something so the stack collapses, pushing something so that it almost falls off but doesn't, kicking soccer ball, pole vault, popping balloons, ripping paper, blowdryhair, throwing knife, playing ice hockey, throwing axe, bending something so that it deforms, yarn spinning, playing cricket, tightrope walking, gymnastics tumbling, playing badminton, archery, pizzatossing, catching or throwing baseball, catching or throwing frisbee, chiseling stone, spinning something that quickly stops spinning, folding napkins, golf, throwing something, nunchucks, opening bottle (not wine), bending something until it breaks, wiping something off of something, playing field hockey, balancebeam, pushing something off of something, throwing snowballs, pouring beer, bungee jumping, drop kicking, pulling two ends of something so that it separates into two pieces, catching or throwing softball, rowing, hammering, letting something roll up a slanted surface so it rolls back down, playing polo

Table 10: Hard Actions List: A comprehensive collection of challenging physical actions in the "Hard" category used to test video models' ability to produce videos that align with physical commonsense

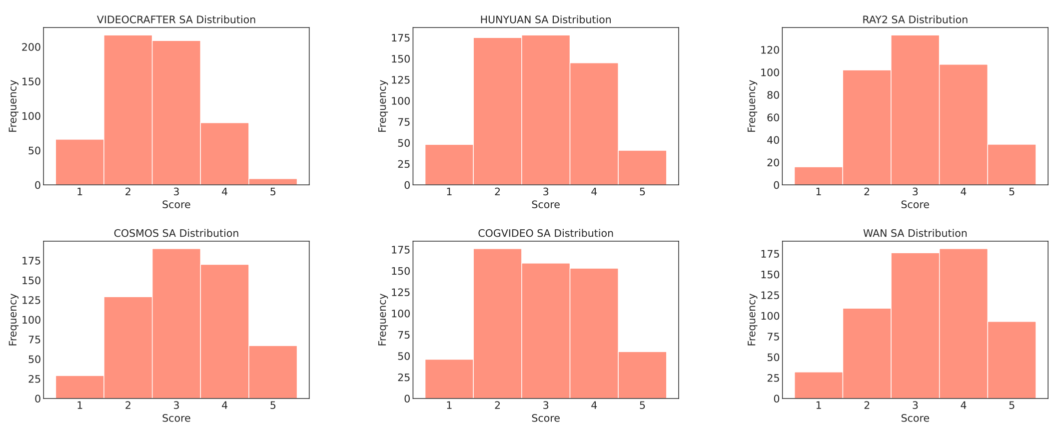


Figure 9: Distribution of semantic adherence scores for various models in VIDEOPHY-2.

Table 11: **Sample prompts in the VIDEOPHY-2.** We present the action and prompt, along with their category—either focusing on physical activities or object interactions. Additionally, we highlight the potential physical principle associated with each prompt.

Action	Prompt	Physical Principle	Category
Canoeing	A person uses a kayak paddle to push their kayak up the bank of a river.	Buoyancy	Physical Activity
Riding Unicycle	A person on a unicycle stops abruptly, putting a foot down to regain balance.	Inertia	Physical Activity
Pushing something so it spins	A chef pushes a pizza spinning on a tray with their hand.	Conversation of Momentum	Object Interaction

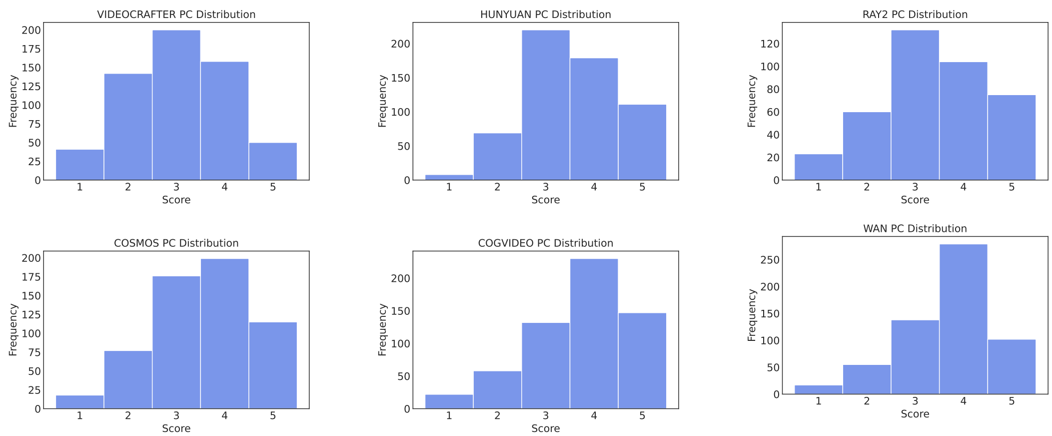


Figure 10: Distribution of physical commonsense scores for various models in VIDEOPHY-2.

Task: Generate 20 realistic, detailed, and diverse video prompts based on a given action. The prompts must focus solely on clear, observable physical actions and interactions between characters, objects, and their environments that can be easily represented visually in a video.
Requirements:
1. **Physical Actions with Direct Outcomes:** Each prompt must clearly describe observable physical actions involving objects (e.g., bows, arrows, targets, or archers) and their direct, visible outcomes (e.g., an arrow striking a target, string tension when pulled).
2. **Exclude Non-Visual Details:** Avoid including sensory or inferred details like sounds, smells, emotions, or mental states (e.g., ”focus,” ”determination,” or ”excitement”).
3. **Avoid Subtle Movements:** Do not include descriptions of small-scale or subtle motions that are hard to detect in standard video playback (e.g., trembling hands, tiny vibrations, or imperceptible shifts).
4. **Concrete, Not Abstract:** Steer clear of poetic, artistic, or abstract descriptions (e.g., “the shimmering flight of an arrow” or “a poised stance”) and instead focus on tangible, visible actions.
5. **Diverse Scenarios:** Use a variety of settings, character types, objects, and equipment to make the prompts diverse and adaptable to different video generation contexts.
6. **Specific Visual Actions:** Center prompts on observable, specific actions such as pulling a bowstring, releasing an arrow, an arrow’s flight path, or its interaction with targets. Highlight the physical interactions between objects (e.g., wood, metal, or other materials) and environments.
Examples:
Good Prompts for the Action “Archery”
”An archer draws the bowstring back to full tension, then releases the arrow, which flies straight and strikes a bullseye on a paper target.”
”A compound bow is fired, and the arrow pierces through a stack of hay bales, stopping halfway through.”
”An archer adjusts their stance, takes aim, and releases an arrow, which embeds itself into a foam target with visible force.”
”A crossbow is loaded with a bolt, cocked, and fired, hitting a glass bottle, which shatters on impact.”
”An arrow is shot toward a wooden target, splinters flying as it embeds deep into the surface.”
Bad Prompts for the Action “Archery”
”The archer feels confident as they aim their arrow at the target.” ## (Describes inferred mental state.)
”The arrow flies silently and gracefully toward the target.” ## (Includes non-visual elements and artistic descriptions.)
”The string vibrates slightly after the arrow is released.” ## (Focuses on subtle, hard-to-detect motion.)
”The archer holds their bow with a poised and elegant stance.” ## (Focuses on posture instead of action.)
Now, Generate up to 20 prompts relevant to the given action: ”” while adhering to the above criteria. Use various objects, environments, and physical interactions to ensure diversity and realism. Output in a python-parsable list of strings, stored in the variable “‘prompts“‘.

Figure 11: LLM Prompt for Generating Realistic Video Captions.

Task Description: Generate simple, clear behaviors/rules for describing visible, real-world physical interactions in a given video scene. These behaviors will be used to determine whether a video aligns with realistic physical interactions.
Key Requirements:
1. **Observation-Centric:** Focus strictly on what is visually observable in the video (e.g., motion, deformation, changes in shape or position). Avoid abstract concepts or invisible factors (e.g., forces, emotions, intentions).
2. **Action-Oriented:** The behaviors should directly describe interactions between materials or objects, not actions or intentions of individuals. The focus should be on the materials' behavior (e.g., how the gauze behaves when wrapped, how it stretches, or conforms), not the motions of the hands or other actors.
3. **Simple and Testable:** The behaviors must be concise, clear, and directly testable from the video. Avoid jargon or unnecessary technical details.
4. **Associated Physical Laws:** Each rule should include the physical law(s) it exemplifies. Use applicable laws from the following list:
"Gravity", "Buoyancy", "Elasticity", "Friction", "Conservation of Mass",
"Reflection", "Refraction", "Interference and Diffraction", "Tyndall Effect"
"Sublimation", "Melting", "Boiling", "Liquefaction"
"Hardness", "Solubility", "Dehydration Properties", "Flame Reaction"
Other laws are acceptable (e.g., Archimedes' Principle), but they must be well-known physical laws. Rules should only be related to visible, observable physical properties of materials, such as shape, deformation, or material properties, and not vague statements that are not strictly physical phenomena.
5. **Guaranteed:** The behaviors must occur in the video as specified from the prompt description. Events that are not guaranteed to occur from the prompt description should not appear in the rules list. Do not add information beyond what is present in the caption.
Example Prompt: *A basketball bounces up and down*
Good Examples of Behaviors:
- The ball is faster at the bottom of the bounce. (Gravity)
- The ball moves up after bouncing off the floor. (Elasticity)
Bad Examples of Behaviors:
- The gravity acts on the ball. (## Not specific enough to test.)
- The ball deforms when it collides with the floor. (## Difficult to test.)
- The ball is caught if it is being dribbled. (## Not guaranteed from the prompt.)
- The ball stops bouncing after a while. (## Not guaranteed; the prompt does not specify video duration.)
Example Prompt: *A nurse applies a bandage around a patient's arm.*
Good Examples of Behaviors:
- The bandage is a flexible and solid material. (Elasticity)
- The bandage roll becomes smaller as it is unrolled. (Conservation of Mass)
Bad Examples of Behaviors:
- The bandage stops the bleeding. (## No bleeding mentioned in the prompt.)
- The bandage secures the wound tightly. (## Tightness is subjective and not visually testable.)
- The bandage is elastic and stretches when pulled. (## Elasticity applies, but not directly related to the video's action.)
Now, generate a suitable list of 3 behaviors for this prompt:
A person shoots an explosive arrow at a metal target.
Output in a Python-comprehensible format with:
"'[("rule1", ["Melting","Gravity"]), ("rule2", ["Conservation of Momentum"]), ..., etc.]"'
Stored in the variable: "'behaviors"'.

Figure 12: LLM Prompt for Generating Physical Rules.

Task Description: Evaluate the alignment between a video and its accompanying caption: *"A badminton shuttlecock is served underhand, traveling across the net and landing within the service box."*
Evaluation Criteria:
1. **Entities and Objects:** Do the objects, entities, and subjects mentioned in the caption appear in the video?
2. **Actions and Events:** Are the actions, events, or interactions described in the caption clearly depicted in the video?
3. **Temporal Consistency:** If the caption describes a sequence or progression of events, does the video follow the same temporal order?
4. **Scene and Context:** Does the overall scene (e.g., background, setting) match the description in the caption?
Instructions for Scoring:
- **1:** No alignment. The video does not match the caption at all (e.g., different objects, events, or scene).
- **2:** Poor alignment. Only a few elements of the caption are depicted, but key objects or events are missing or incorrect.
- **3:** Moderate alignment. The video matches the caption partially, but there are inconsistencies or omissions.
- **4:** Good alignment. Most elements of the caption are depicted correctly in the video, with minor issues.
- **5:** Perfect alignment. The video fully adheres to the caption with no inconsistencies.
Example Prompt: *"A badminton shuttlecock is served underhand, traveling across the net and landing within the service box."*
Example Responses:
Score: 3 **Explanation:** The shuttlecock is present, and an underhand serve is performed. However, the landing position is unclear, and the trajectory is partially obstructed.

Figure 13: Gemini Prompt for Evaluating Semantic Adherence.

Original Caption	Upsampled Caption
A person uses nunchucks to break a stack of wooden blocks.	In a dynamic display of martial arts prowess, a skilled practitioner wields a pair of nunchucks, their hands clad in black gloves that enhance grip and safety. The scene unfolds in a dimly lit, industrial setting, where a stack of wooden blocks, meticulously arranged in a pyramid, awaits the test of strength. The camera captures the moment with a static focus, highlighting the nunchucks' fluid motion as they swing through the air, their polished wooden surfaces glinting under the soft, ambient light.
A player throws a softball sidearm, and the ball spins as it travels through the air.	In a sun-drenched outdoor sports arena, a dynamic softball game unfolds, captured with cinematic precision. The camera, positioned at a low angle, focuses on a player in a sleek black jersey, their arm poised in a powerful sidearm throwing motion. The softball, a vibrant white against the lush green field, spins rapidly as it leaves their hand, tracing a graceful arc through the air.
A small rock tumbles down a steep, rocky hillside.	In a breathtaking display of nature's raw power, a small, dark rock careens down a steep, rugged hillside, its descent punctuated by a cascade of displaced soil and smaller stones. The camera, positioned at a static angle, captures the dynamic interplay of gravity and friction as the rock bounces and rolls, kicking up a cloud of dust that dances in the sunlight.
An athlete throws a hammer, with the hammer head clearly visible in rotation.	In a breathtaking outdoor setting, a skilled athlete stands poised, clad in a sleek black and white athletic uniform, their focus unwavering as they prepare to hurl a gleaming hammer. The hammer rotates gracefully, its head spinning in a mesmerizing blur against the backdrop of a clear blue sky.
A person on a unicycle stops abruptly, putting a foot down to regain balance.	In a serene outdoor setting, a solitary figure clad in a vibrant red jacket and black pants gracefully navigates a narrow, winding path on a sleek black unicycle. The camera captures the scene with a static shot, allowing the viewer to fully appreciate the rider's skill and balance as they glide effortlessly forward.

Table 12: Example of upsampled captions given to video generation models

Task Description: Evaluate whether the video follows physical commonsense. This judgment is based solely on the video itself and does not depend on the caption.
Evaluation Criteria:
1. **Object Behavior:** Do objects behave according to their expected physical properties (e.g., rigid objects do not deform unnaturally, fluids flow naturally)?
2. **Motion and Forces:** Are motions and forces depicted in the video consistent with real-world physics (e.g., gravity, inertia, conservation of momentum)?
3. **Interactions:** Do objects interact with each other and their environment in a plausible manner (e.g., no unnatural penetration, appropriate reactions on impact)?
4. **Consistency Over Time:** Does the video maintain consistency across frames without abrupt, unexplainable changes in object behavior or motion?
Instructions for Scoring:
- **1:** No adherence to physical commonsense. The video contains numerous violations of fundamental physical laws.
- **2:** Poor adherence. Some elements follow physics, but major violations are present.
- **3:** Moderate adherence. The video follows physics for the most part but contains noticeable inconsistencies.
- **4:** Good adherence. Most elements in the video follow physical laws, with only minor issues.
- **5:** Perfect adherence. The video demonstrates a strong understanding of physical commonsense with no violations.
Example Responses:
Score: 2 **Explanation:** The ball's motion is inconsistent with gravity; it hovers momentarily before falling. Additionally, object interactions lack expected momentum transfer, suggesting physics inconsistencies.

Figure 14: Gemini Prompt for Evaluating Physical Commonsense.

Semantic Adherence:

Given: V (Video), **T** (Caption)

Instruction (I): *[V] Does this video match the description: "[**T**]"? Please rate the video on a scale from 1 to 5, where 5 indicates a perfect match and 1 indicates no relevance.*

Response (R): *1, 2, 3, 4, or 5*

Figure 15: Template for assessing semantic adherence using a multi-modal VLM.

Physical Commonsense:

Given: V (Video)

Instruction (I): *[V] Does this video adhere to physical laws? Rate the video on a scale from 1 to 5, where 5 means full compliance and 1 means significant violations.*

Response (R): *1, 2, 3, 4, or 5*

Figure 16: Template for assessing physical commonsense in video-based evaluation.

Rule Validation:

Given: V (Video), **R** (Physical Rule)

Instruction (I): *[V] Does the video follow the physical rule: "[**R**]"? Choose 0 if not, 1 if valid, or 2 if indeterminate.*

Response (R): *0, 1, or 2*

Figure 17: Template for validating specific physical rules in a video.

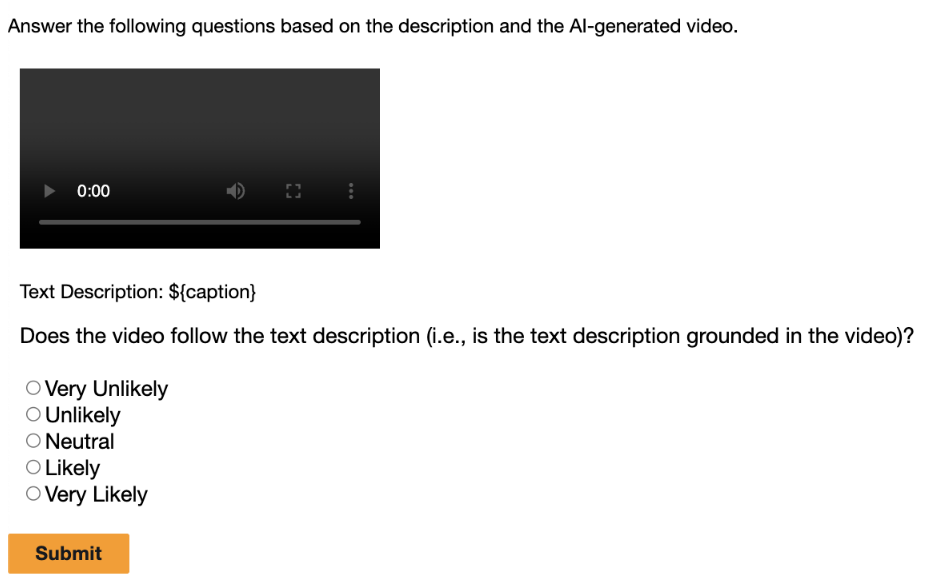
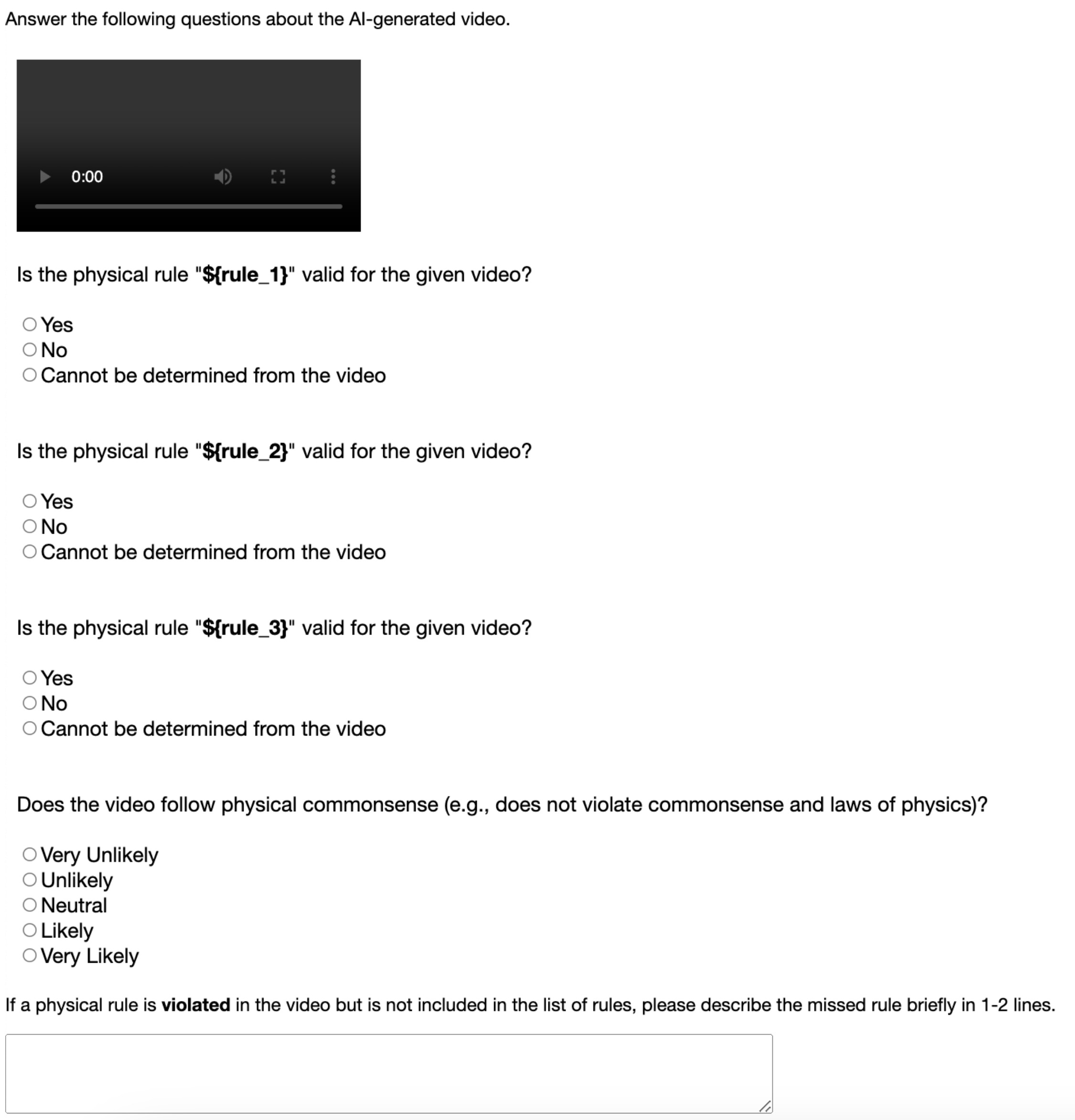


Figure 18: The screenshot of the human annotation interface for semantic adherence task.


Answer the following questions about the AI-generated video.

0:00

Is the physical rule "**${rule_1}**" valid for the given video?

- Yes
- No
- Cannot be determined from the video

Is the physical rule "**${rule_2}**" valid for the given video?

- Yes
- No
- Cannot be determined from the video

Is the physical rule "**${rule_3}**" valid for the given video?

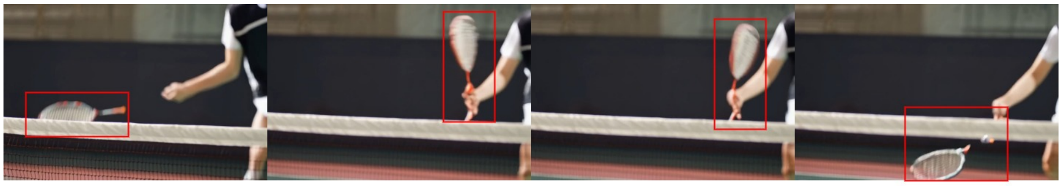
- Yes
- No
- Cannot be determined from the video

Does the video follow physical commonsense (e.g., does not violate commonsense and laws of physics)?

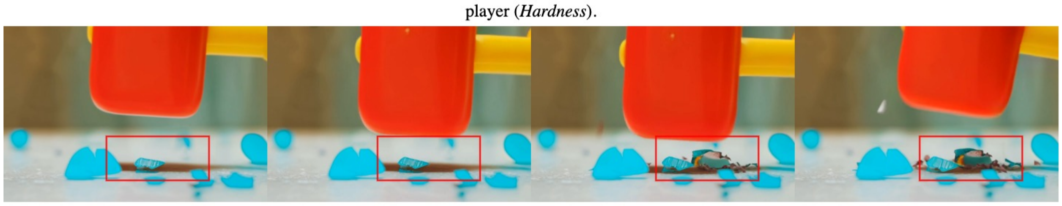
- Very Unlikely
- Unlikely
- Neutral
- Likely
- Very Likely

If a physical rule is **violated** in the video but is not included in the list of rules, please describe the missed rule briefly in 1-2 lines.

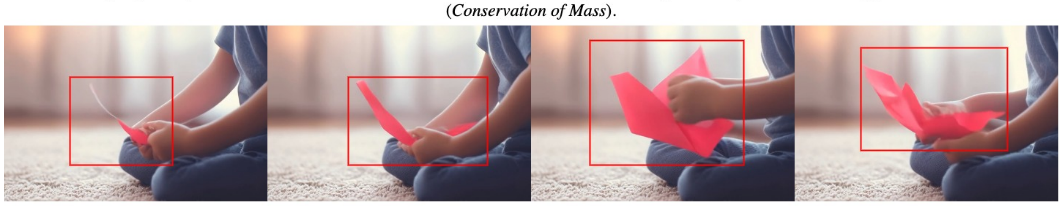
Figure 19: The screenshot of the human annotation interface for physical rule and commonsense judgment tasks.



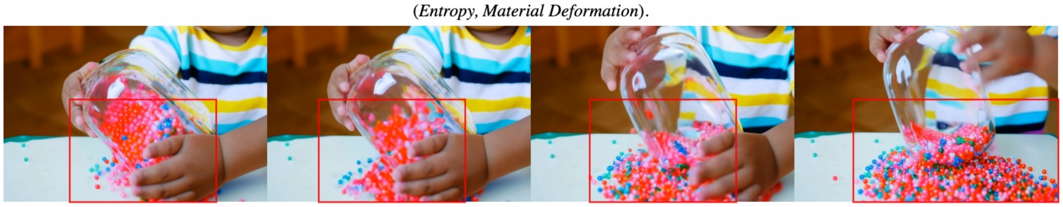
(a) **Caption:** A badminton shuttlecock is served underhand, traveling across the net and landing within the service box.
Human-judged Physical Violations: The badminton racket should maintain its rigid shape and should only change directions when held by the player (*Hardness*).



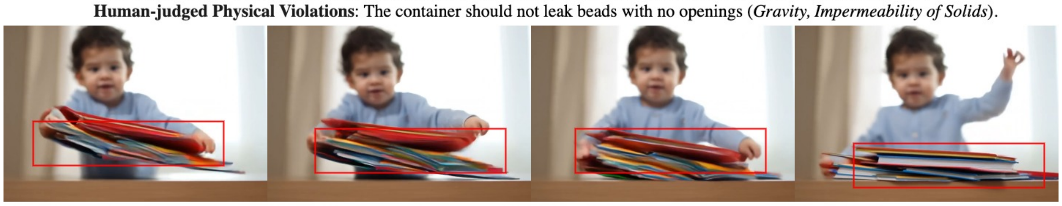
(b) **Caption:** A child's toy hammer smashes a small plastic egg, breaking it open.
Human-judged Physical Violations: The mass and amount of shards of egg should stay constant, no mass should appear out of nowhere (*Conservation of Mass*).



(c) **Caption:** A child folds a piece of origami paper into a simple crane, with visible creases appearing.
Human-judged Physical Violations: Creases should not spontaneously form in the paper without an external force or proper folding sequence (*Entropy, Material Deformation*).

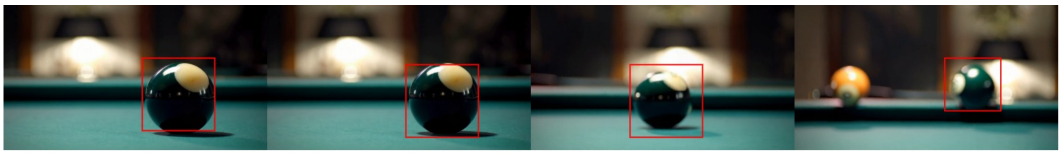


(d) **Caption:** A child pours colorful beads from a plastic container into a glass jar until they overflow, scattering on the floor.
Human-judged Physical Violations: The container should not leak beads with no openings (*Gravity, Impermeability of Solids*).

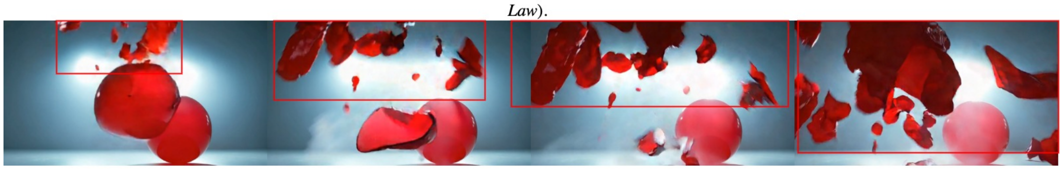


(e) **Caption:** A child pushes a stack of books off a desk; the books fall in a chaotic pile.
Human-judged Physical Violations: The amount and color of the books should not change as a result of setting them down (*Conservation of Mass*).

Figure 20: Examples of physically unlikely video generations from CogVideoX-5B. Each case demonstrates violations of fundamental physical laws.

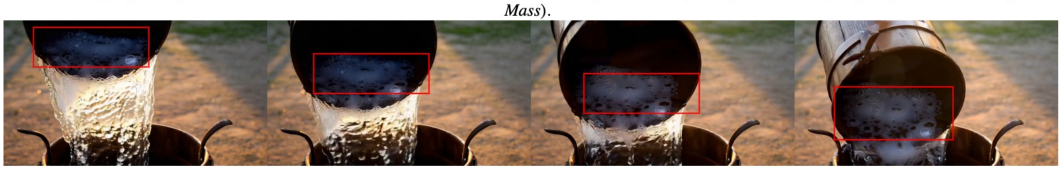


(a) **Caption:** A billiard ball is struck and rolls across a felt-covered pool table, hitting another ball.
Human-judged Physical Violations: The billiard ball cannot move and jump into the air without an external force acting on it (*Newton's First Law*).

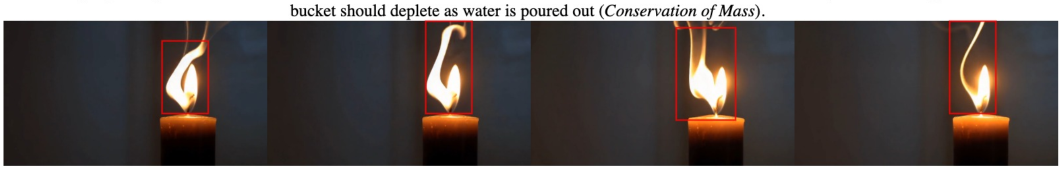


(b) **Caption**: A balloon is popped by being over-inflated, causing it to burst with a loud noise and rapid deflation.
Human-judged Physical Violations: The total volume of rubber should remain consistent (*Conservation of Mass*).

(c) **Caption**: A bowling ball rolls down a polished wooden lane, hitting the pins at the end.
Human-judged Physical Violations: The surfer and the board cannot be submerged at the beginning of the wave (*Buoyancy, Conservation of Mass*).

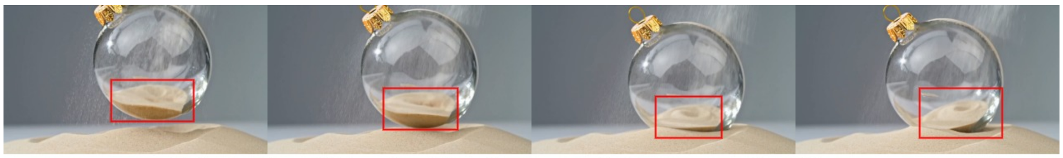


(d) **Caption**: A bucket filled with water tips over, and the water splashes and spills onto the ground.
Human-judged Physical Violations: Bubbles on the surface of the water should flow outwards along with the water (*Inertia, Gravity*). Water in the bucket should deplete as water is poured out (*Conservation of Mass*).



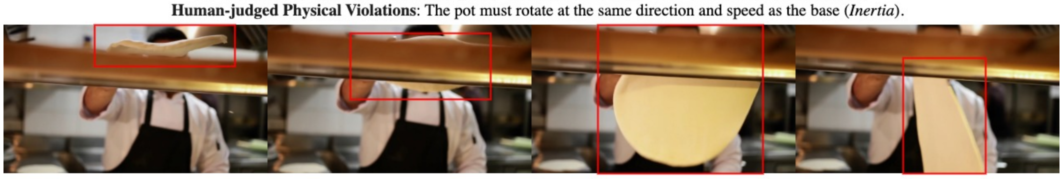
(e) **Caption**: A burning candle produces a small amount of smoke near its wick.
Human-judged Physical Violations: The flame should be affected by the same airflow that is affecting the smoke (*Fluid Dynamics*). The smoke should be less illuminated than the fire (*Combustion*).

Figure 21: Examples of physically unlikely video generations from Cosmos. Each case demonstrates violations of fundamental physical laws.



(a) **Caption**: Fine sand is directed at a delicate glass ornament with light pressure, revealing a hidden pattern.
Human-judged Physical Violations: The ornament should not collect sand internally if there are no openings, and the sand should not move inside a stationary ornament (*Gravity, Impermeability of Solids*).

(b) **Caption**: A potter's wheel is lightly poked with a tool, causing a small wobble in the rotating clay.
Human-judged Physical Violations: The pot must rotate at the same direction and speed as the base (*Inertia*).



(c) **Caption**: A pizza chef uses a pizza peel to toss and catch a pizza, then slides the pizza onto a baking sheet.
Human-judged Physical Violations: The dough's elasticity and hardness should stay constant and not become too hard or soft throughout the video (*Elasticity, Hardness, Gravity*).

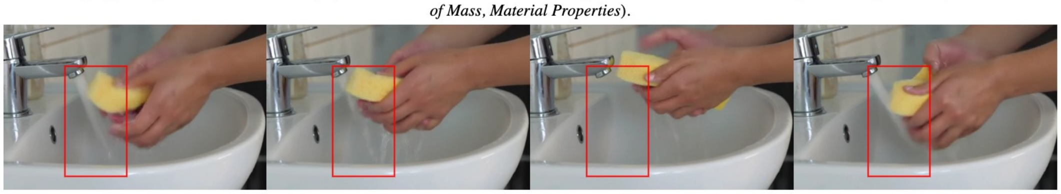
(e) **Caption**: Time-lapse of a pencil being sharpened with sandpaper; the pencil point becomes increasingly sharp.
Human-judged Physical Violations: The pencil should maintain its general shape and the mass that it loses should be visible (*Hardness, Conservation of Mass*).

(d) **Caption**: Using a knife, a baker cuts a loaf of pre-shaped dough into smaller rolls, making visible cuts.
Human-judged Physical Violations: The knife should protrude out of the roll given its length, and should leave a mark on the dough once removed (*Conservation of Mass, Elasticity, Friction*).

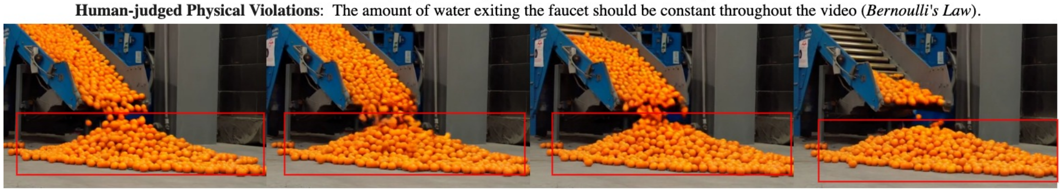
Figure 22: Examples of physically unlikely video generations from Ray2. Each case demonstrates violations of fundamental physical laws.

(a) **Caption**: Someone throws a water balloon at a car windshield; the balloon bursts, leaving water residue.
Human-judged Physical Violations: The balloon should collapse and lose its shape upon bursting, with its volume decreasing as the water is expelled (*Conservation of Mass, Material Deformation*).

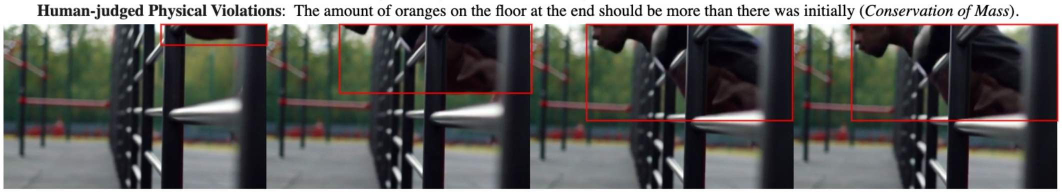
(b) **Caption**: A worker uses a pneumatic chisel to break apart a large chunk of concrete, sending debris flying.
Human-judged Physical Violations: A large piece of concrete cannot spontaneously form from the dust and debris generated by the drill (*Conservation of Mass, Material Properties*).



(c) **Caption**: A wet sponge is wrung out in a sink, leaving behind a puddle of water.
Human-judged Physical Violations: The amount of water exiting the faucet should be constant throughout the video (*Bernoulli's Law*).

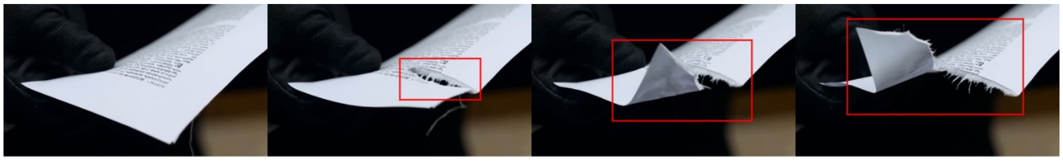


(d) **Caption**: A bag of oranges falls from a conveyor belt and scatters on the floor.
Human-judged Physical Violations: The amount of oranges on the floor at the end should be more than there was initially (*Conservation of Mass*).

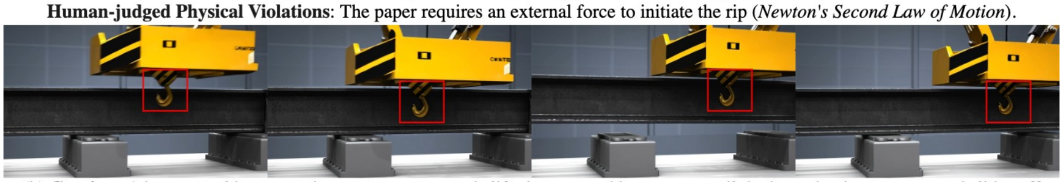


(e) **Caption**: Parallel bars are shown from a side view with an athlete performing a series of dips.
Human-judged Physical Violations: The athlete's body must make contact with the bars during dips and cannot pass through them (*Impenetrability of Matter, Conservation of Momentum*).

Figure 23: Examples of physically unlikely video generations from Hunyuan. Each case demonstrates violations of fundamental physical laws.



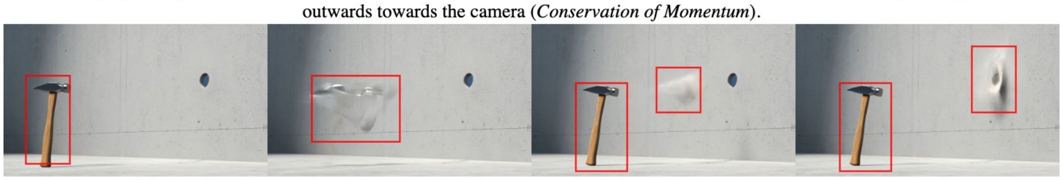
(a) **Caption:** A hand rips a sheet of printer paper in half, creating a jagged tear.
Human-judged Physical Violations: The paper requires an external force to initiate the rip (*Newton's Second Law of Motion*).



(b) **Caption:** A heavy steel beam, resting on two supports, is lifted at one end by a crane until the beam begins to rotate and slides off the supports.
Human-judged Physical Violations: The beam should only be lifted as a result of an upward force exerted by the crane hook that is hooked onto the beam (*Inertia*).

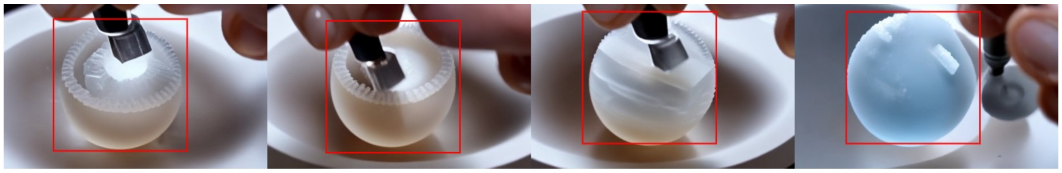
(c) **Caption:** A large stone rolls down a hillside, leaving a visible path in the dirt.
Human-judged Physical Violations: The boulder cannot ascend a slope without an external force applied, and the round shape of the rock indicates that it should be rolling (*Gravity, Friction*).

(d) **Caption:** A bowling ball rolls down a polished wooden lane, hitting the pins at the end.
Human-judged Physical Violations: The ball starts in front of the pins and should not be able to collide with the pins to push them outwards towards the camera (*Conservation of Momentum*).

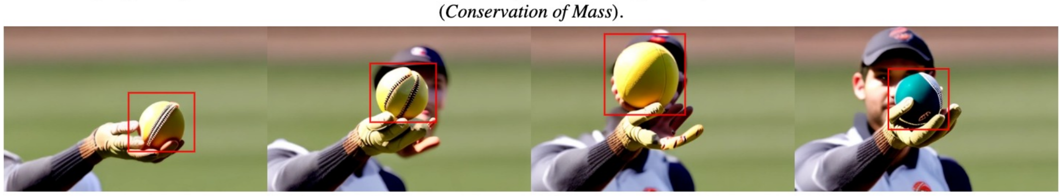


(e) **Caption:** A hammer, thrown with considerable force, bounces off a concrete wall, leaving a visible mark.
Human-judged Physical Violations: There needs to be an external force to push the hammer initially, the hammer should not disappear, and there should not be material ejected from the hammer out of nowhere (*Inertia, Material Properties, Conservation of Mass*).

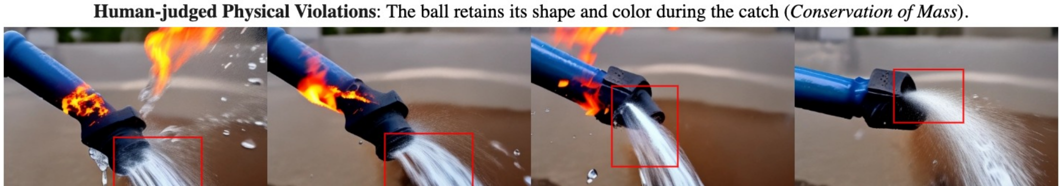
Figure 24: Examples of physically unlikely video generations from Wan2.1. Each case demonstrates violations of fundamental physical laws.



(a) **Caption:** A person carves an ice cube into a perfect sphere using a specialized rotary tool and a steady hand.
Human-judged Physical Violations: The tool maintains its shape while carving the ice, and the ices should not increase in volume (*Conservation of Mass*).

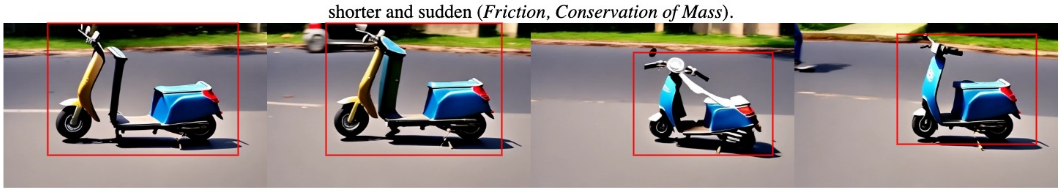


(b) **Caption:** A fielder catches a cricket ball holding it securely in their glove.
Human-judged Physical Violations: The ball retains its shape and color during the catch (*Conservation of Mass*).



(c) **Caption:** A stream of water from a handheld nozzle hits a small oil fire extinguishing the blaze.
Human-judged Physical Violations: The fire should lead a mark and needs to be put out by something, and water should exit the hose at only the nozzle (*Impermeability of Solids, Combustion)*

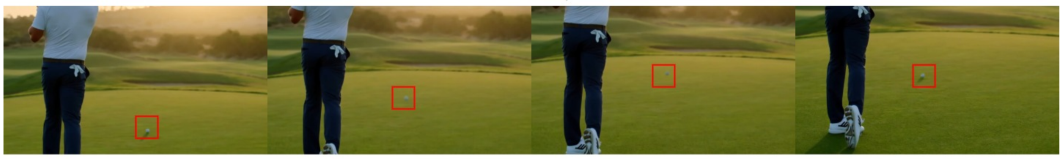
(d) **Caption:** A speeding car crashes into a brick wall crumpling the front end and stopping abruptly.
Human-judged Physical Violations: The car cannot shrink or flip upon impact with the wall, and impact duration should be way shorter and sudden (*Friction, Conservation of Mass*).



(e) **Caption:** A scooter collides with a trash can the scooter tilting to one side before stopping.
Human-judged Physical Violations: The scooter cannot change its model or color spontaneously when in motion (*Conservation of Mass*).

Figure 25: Examples of physically unlikely video generations from VideoCrafter2. Each case demonstrates violations of fundamental physical laws.

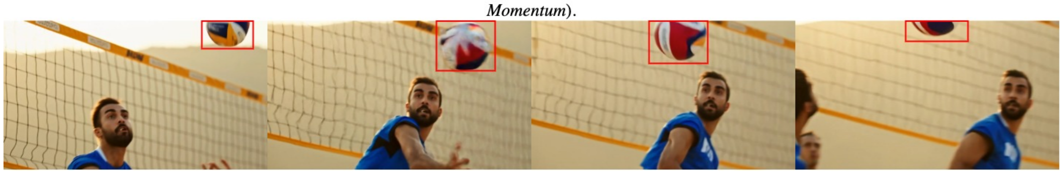
(a) **Model**: Ray2
Caption: A person throws a yo-yo, and it strikes a rubber ball, causing the ball to move.
Human-judged Physical Violations: The red object should only move as a result of a direct impulse from the paddles. (*Conservation of Momentum*).



(b) **Model**: Ray2
Caption: A golf club strikes a golf ball, sending it rolling across a manicured green toward the hole.
Human-judged Physical Violations: The golf ball should only move after being struck by the golf club (*Conservation of Momentum*).

(c) **Model**: Cosmos-1.0
Caption: An object shaped like a globe is poked, causing it to spin on its axis, revealing different countries.
Human-judged Physical Violations: The globe should only spin and move once it is poked (*Conservation of Momentum*).

(d) **Model**: Cosmos-1.0
Caption: A player throws a sidearm pass; the football travels on a sideways trajectory before being caught.
Human-judged Physical Violations: The ball should follow the trajectory dictated by the throw to the right, not upwards (*Conservation of Momentum*).



(e) **Model**: Ray2
Caption: An outside hitter hits a line shot, the ball bouncing sharply off the side line.
Human-judged Physical Violations: The ball should only change direction after colliding with anothe robject (*Conservation of Momentum*).

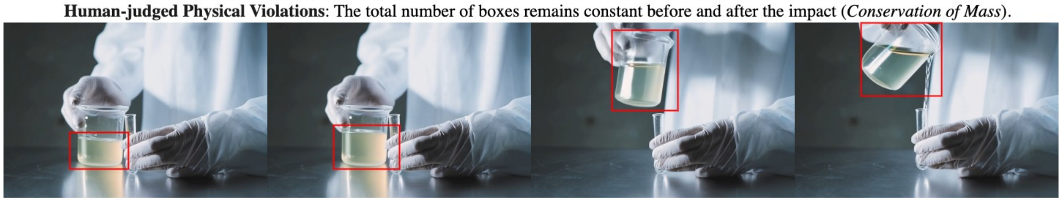
Figure 26: Examples of physically unlikely video generations where the physical law of conservation of momentum is violated

(a) **Model**: Cosmos-1.0
Caption: A pizza dough is tossed up, its edges become thinner as it rotates in the air, and it's caught by a second person.
Human-judged Physical Violations: The pizza dough maintains its volume while stretching during the toss (*Conservation of Mass*).

(b) **Model**: Cosmos-1.0
Caption: A squash player serves the ball, the ball striking the front wall with a visible impact.
Human-judged Physical Violations: The player's left hand should not hold an object that disappears (*Conservation of Mass*).

(c) **Model**: Cosmos-1.0
Caption: A person rips a playing card in half, separating the two halves cleanly.
Human-judged Physical Violations: The total amount of paper should remain constant and not appear out of nowhere (*Conservation of Mass*).

(d) **Model**: Hunyuan
Caption: A car crashes into a stack of cardboard boxes, sending the boxes flying in all directions.
Human-judged Physical Violations: The total number of boxes remains constant before and after the impact (*Conservation of Mass*).



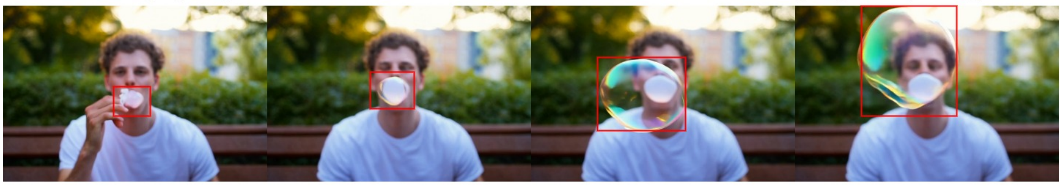
(e) **Model**: CogvideoX-5B
Caption: A chemist pours a clear liquid from a beaker into a test tube, carefully avoiding spills.
Human-judged Physical Violations: The liquid level does not rise in the beaker during pouring and the height of the beaker should not increase (*Conservation of Mass*).

Figure 27: Examples of physically unlikely video generations where the physical law of conservation of mass is violated

(a) **Model**: Cosmos-1.0

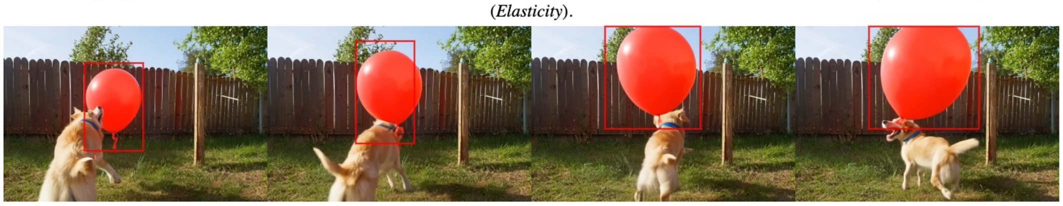
Caption: A rider sits on the mechanical bull, which starts to buck gently; then, the intensity gradually increases.

Human-judged Physical Violations: The bull's metal horns should not deform nor return to their original positions (*Elasticity*).



(b) **Caption**: A person blows a bubble gum bubble while sitting on a park bench, the bubble floating away.

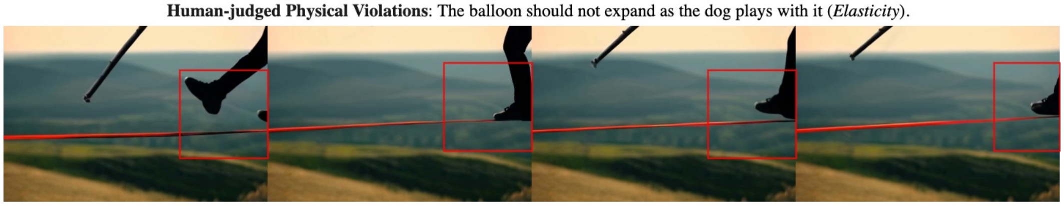
Human-judged Physical Violations: The bubble should not be able to stretch so think that it becomes as thin and transparent as a soap bubble (*Elasticity*).



(c) **Model**: CogvideoX-5B

Caption: A dog playfully bats at a balloon, causing it to pop against a nearby fence.

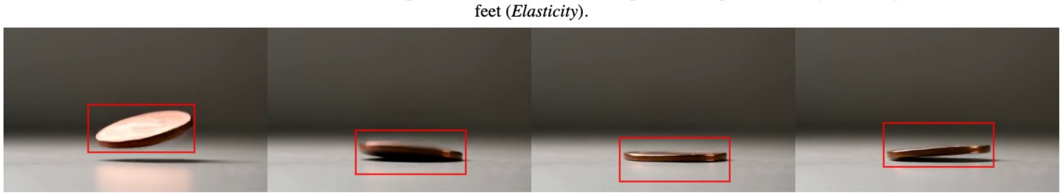
Human-judged Physical Violations: The balloon should not expand as the dog plays with it (*Elasticity*).



(d) **Model**: CogvideoX-5B

Caption: A tightrope walker, with their balance pole extended to the side, uses it to brace against a sudden gust of wind.

Human-judged Physical Violations: The rope should bend or deform in response to the person shifting their weight with their feet (*Elasticity*).



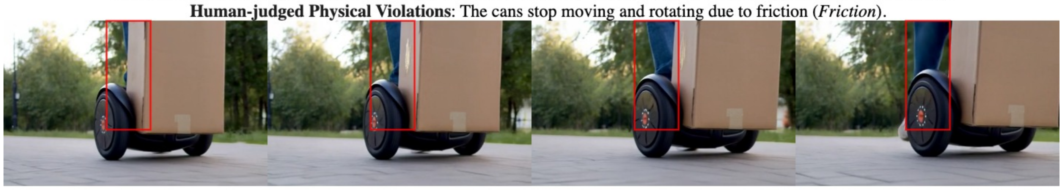
(e) **Model**: Hunyuan

Caption: A coin spins on a flat surface, coming to rest on heads.

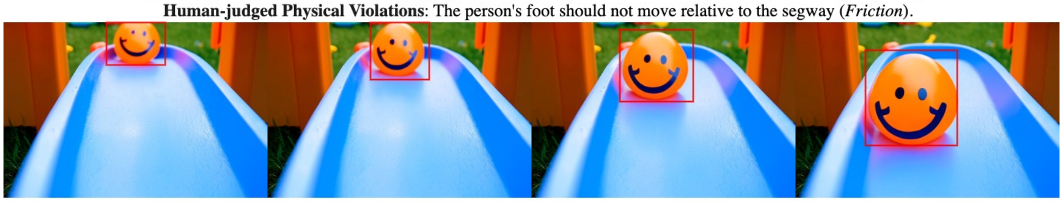
Human-judged Physical Violations: The coin does not bounce off the surface after coming to rest (*Elasticity*).

Figure 28: Examples of physically unlikely video generations where the physical law of elasticity is violated

(a) **Model**: Ray2
Caption: A small child pokes a stack of empty soda cans with a toy car, causing the cans to topple like a chain reaction.
Human-judged Physical Violations: The cans stop moving and rotating due to friction (*Friction*).

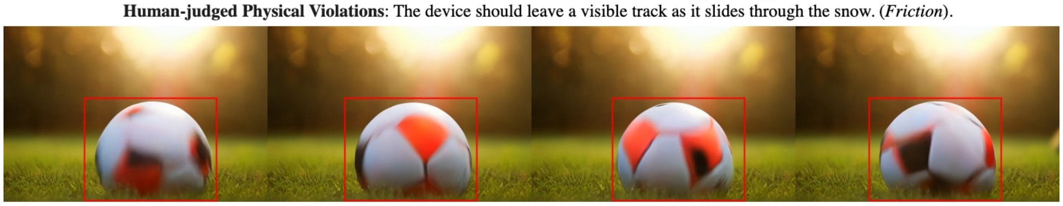


(b) **Model**: Hunyuan
Caption: A package is secured to a Segway's platform; the Segway travels a short distance, delivering the package.
Human-judged Physical Violations: The person's foot should not move relative to the segway (*Friction*).



(c) **Model**: CogvideoX-5B
Caption: A large, orange plastic egg is rolled up a plastic slide, then rolls down, ending at the bottom.
Human-judged Physical Violations: The ball should roll down the slide with rotational motion. (*Friction*).

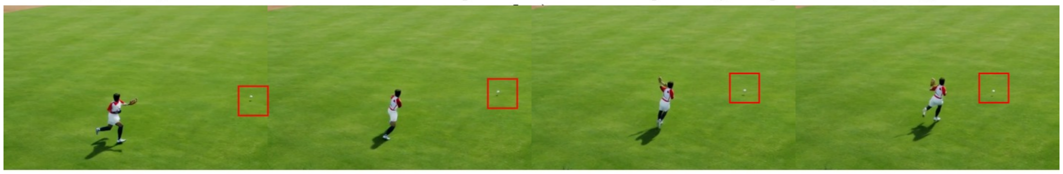
(d) **Model**: Cosmos-1.0
Caption: A robotic arm uses a shovel to clear snow from a large parking lot.
Human-judged Physical Violations: The device should leave a visible track as it slides through the snow. (*Friction*).



(e) **Model**: CogvideoX-5B
Caption: A soccer ball is kicked and hits a person's leg, deflecting at a different angle.
Human-judged Physical Violations:The ball should experience an acceleration in speed as it spins in with the grass (*Friction*).

Figure 29: Examples of physically unlikely video generations where the physical law of friction is violated

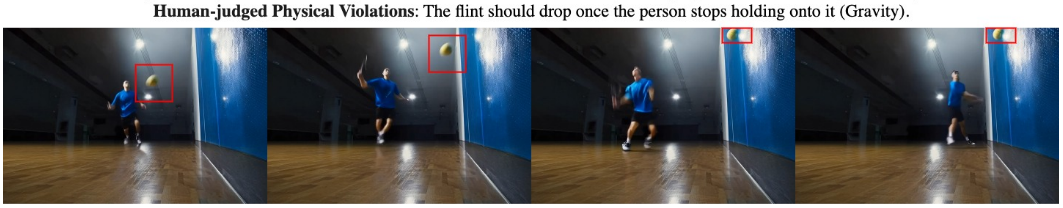
(a) **Model**: CogvideoX-5B
Caption: A large stone rolls down a hillside, leaving a visible path in the dirt.
Human-judged Physical Violations: Rock should not exit its place in the dirt without experiencing an impulse out of it (*Gravity, Newton's First*



(b) **Model**: Hunyuan
Caption: A player hits a grounder, the ball rolling smoothly across the perfectly manicured infield.
Human-judged Physical Violations: The ball should not be able to levitate above the ground (Gravity).

(c) **Model**: Hunyuan
Caption: A discus is thrown from a raised platform, landing sharply in the soft sand of a beach, creating a visible impact crater.
Human-judged Physical Violations: The disc should not be able to leave the crater from rest (Gravity).

(d) **Model**: Cosmos-1.0
Caption: A person uses a variety of tools (hammerstone, antler, bone) to knap a piece of flint into a specific shape.
Human-judged Physical Violations: The flint should drop once the person stops holding onto it (Gravity).



(e) **Model**: CogvideoX-5B
Caption: A player uses a backhand shot to send a racquetball into the corner, the ball rebounding off two walls.
Human-judged Physical Violations: he balls follow the laws of motion and cannot change direction without an external force acting upon them (Gravity).

Figure 30: Examples of physically unlikely video generations where the physical law of gravity is violated

(a) **Model**: Cosmos-1.0
Caption: A bodysurfer dives headfirst into a wave, emerging from the whitewater several seconds later.
Human-judged Physical Violations: The surfer should not be able to stay submerged during the first part of the video due to the buoyancy from their board (*Buoyancy*).

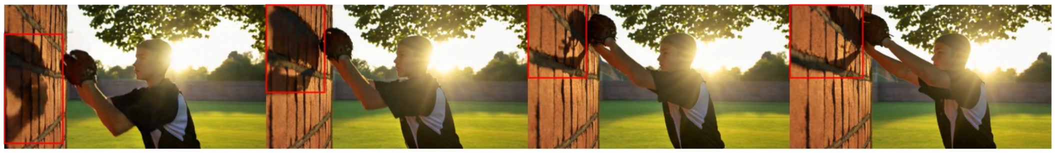
(b) **Model**: Cosmos-1.0
Caption: A kayaker lifts their paddle out of the water, showing the surface tension of the water clinging to the blade.
Human-judged Physical Violations: The woman is too deep in the water such that the kayak should be sinking (*Buoyancy*).

(c) **Model**: Cosmos-1.0
Caption: A person carries a heavy bucket while wading through chest-deep water.
Human-judged Physical Violations: The bucket should not have another liquid inside of it and if it did have a denser liquid it would be much harder to carry (*Buoyancy*).

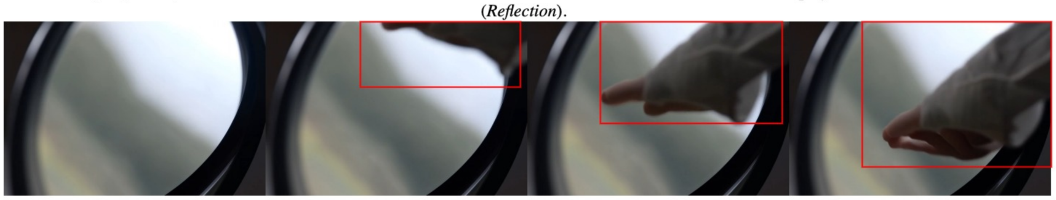
(d) **Model**: Cosmos-1.0
Caption: A surfer falls off their board, colliding briefly with another surfer, before swimming away.
Human-judged Physical Violations: The surfer's legs should remain above the water surface when not on the board (*Buoyancy*).

(e) **Model**: Ray2
Caption: A surfer dives underwater to avoid a breaking wave, resurfacing seconds later.
Human-judged Physical Violations: The surfer should create a spash when entering the water and reappear shortly after (*Buoyancy*).

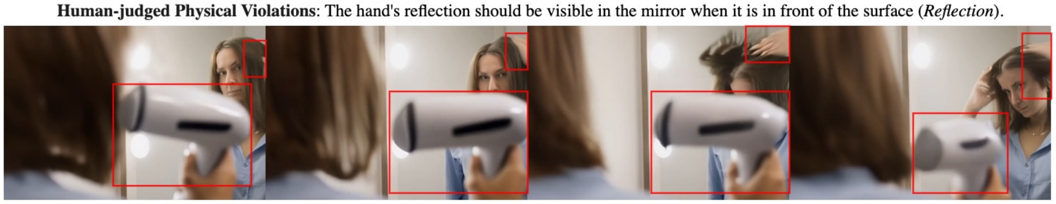
Figure 31: Examples of physically unlikely video generations where the physical law of Buoyancy is violated



(a) **Model**: Cosmos-1.0
Caption: A player throws a softball against a brick wall; the ball rebounds at a visible angle.
Human-judged Physical Violations: The shadow on the brick wall should exhibit the outline of both of the player's arms with no distortions (*Reflection*).

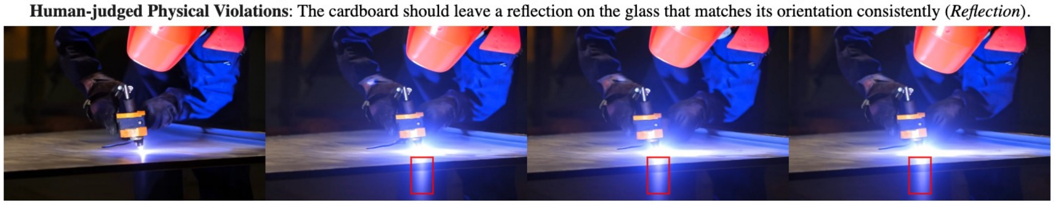


(b) **Model**: CogvideoX-5B
Caption: A large, circular mirror is slightly poked, causing a minor shift in its reflected image, though it does not spin.
Human-judged Physical Violations: The hand's reflection should be visible in the mirror when it is in front of the surface (*Reflection*).



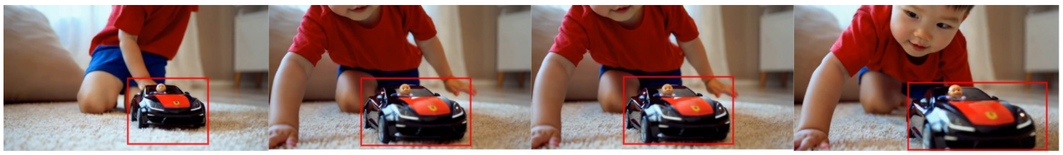
(c) **Model**: CogvideoX-5B
Caption: A person uses a low heat setting on the hairdryer to gently dry their fine hair.
Human-judged Physical Violations: The hairdryer should be visible in the mirror's reflection (*Reflection*).

(d) **Model**: CogvideoX-5B
Caption: A small, rectangular piece of cardboard is placed on a sloped glass surface; it slides down slowly, leaving no visible trace.
Human-judged Physical Violations: The cardboard should leave a reflection on the glass that matches its orientation consistently (*Reflection*).



(e) **Model**: CogvideoX-5B
Caption: A welder uses a plasma cutter to cut a steel sheet, producing a bright arc and a clean cut edge.
Human-judged Physical Violations: Light should not pass through the steel sheet and reflect off of the air into the camera (*Reflection*).

Figure 32: Examples of physically unlikely video generations where the physical law of reflection is violated



(a) **Model**: Cosmos-1.0

Caption: A child pushes a toy car across a carpeted floor; the car rolls a short distance before stopping.

Human-judged Physical Violations: The car should not move after it stops, unless pushed again by an external force (*Inertia*).

(b) **Model**: Ray2

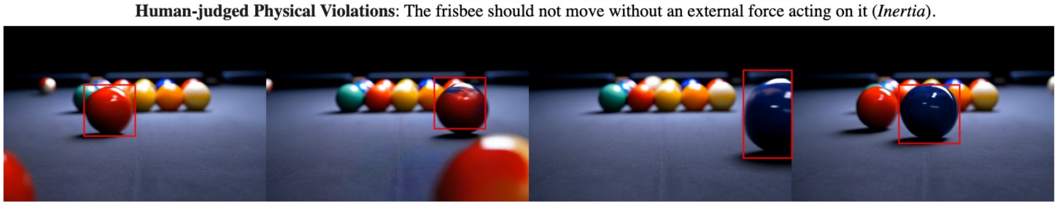
Caption: Concrete is poured from a wheelbarrow into a form, leveling the surface with a trowel.

Human-judged Physical Violations: The wheelbarrow should not move since the man is not touching it (*Inertia*).

(c) **Model**: CogvideoX-5B

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

Human-judged Physical Violations: The frisbee should not move without an external force acting on it (*Inertia*).



(d) **Model**: CogvideoX-5B

Caption: Multiple balls are grouped near a pocket; a precise shot pockets two of the balls.

Human-judged Physical Violations: The frisbee should not move without an external force acting on it (*Inertia*).

(e) **Model**: Hunyuan

Caption: Multiple balls are grouped near a pocket; a precise shot pockets two of the balls.

Human-judged Physical Violations: The bearing should change direction only due to an applied force or interaction with another object and should only stop rotation as a result of an external torque (*Inertia*).

Figure 33: Examples of physically unlikely video generations where the physical law of inertia is violated