VIDEOPHY: Evaluating Physical Commonsense for Video Generation

Hritik Bansal[*1] **Zongyu Lin**[*1] **Tianyi Xie**[†1] **Zeshun Zong**[†1] **Michal Yarom**[‡2]
Yonatan Bitton[‡2] **Chenfanfu Jiang**[1] **Yizhou Sun**[1] **Kai-Wei Chang**[1] **Aditya Grover**[1]

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

https://videophy.github.io/

Abstract

Recent advances in internet-scale video data pretraining have led to the development of text-to-video generative models that can create high-quality videos across a broad range of visual concepts and styles. Due to their ability to synthesize realistic motions and render complex objects, these generative models have the potential to become general-purpose simulators of the physical world. However, it is unclear how far we are from this goal with the existing text-to-video generative models. To this end, we present VIDEOPHY, a benchmark designed to assess whether the generated videos follow physical commonsense for real-world activities (e.g. marbles will roll down when placed on a slanted surface). Specifically, we curate a list of 688 captions that involve interactions between various material types in the physical world (e.g., solid-solid, solid-fluid, fluid-fluid). We then generate videos conditioned on these captions from diverse state-of-the-art text-to-video generative models, including open models (e.g., VideoCrafter2) and closed models (e.g., Lumiere from Google, Pika). Further, our human evaluation reveals that the existing models severely lack the ability to generate videos adhering to the given text prompts, while also lack physical commonsense. Specifically, the best performing model, Pika, generates videos that adhere to the caption and physical laws for only 19.7% of the instances. VIDEOPHY thus highlights that the video generative models are far from accurately simulating the physical world. Finally, we also supplement the dataset with an auto-evaluator, VIDEOCON-PHYSICS, to assess semantic adherence and physical commonsense at scale.

1 Introduction

The ability to synthesize high-quality videos for a broad range of visual concepts and styles is a long-standing goal of generative modeling [1]. In this regard, recent advancements in pretraining on internet-scale video data [2, 89, 84, 82, 21] have led to the development of various text-to-video (T2V) generative models such as Sora [46] that can generate photo-realistic videos conditioned on a text prompt [7, 81, 20, 55, 71, 13, 37]. Specifically, these models can generate complex scenes (e.g., 'busy street in Japan') and realistic motions (e.g., 'running', 'pouring'), making them amenable for understanding and simulating the physical world. Recent efforts [23, 17] have further utilized text-guided video generation to train agents that can act, plan, and solve goals in the real world. In spite of the strong physical motivations of these works, it remains unclear *how well the generated videos from T2V models adhere to the laws of physics*.

[*†‡] Equal Contribution.

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

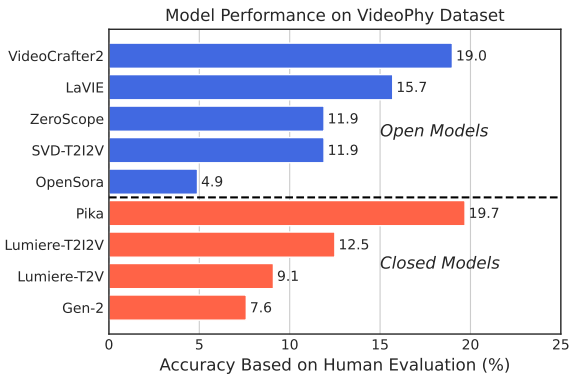


Figure 1: **Model performance on the VIDEOPHY dataset using human evaluation.** We assess the physical commonsense and semantic adherence to the conditioning caption in the generated videos. We find that Pika (closed model) and VideoCrafter2 (open model) can generate videos that follow the caption and physics laws for 19.7% and 19% of the prompts, respectively. This indicates that the existing models are quite far from being general-purpose physical world simulators.

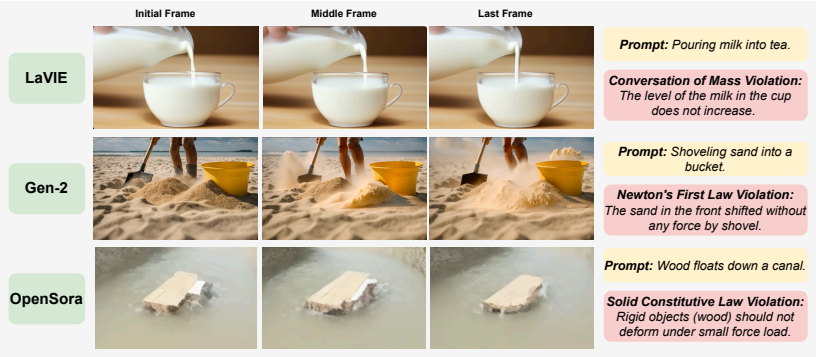


Figure 2: **Illustration of poor physical commonsense by various T2V generative models.** Here, we show that the generated videos can violate a diverse range of physics laws such as conversation of mass, Newton's first law, and solid constitutive laws. In VIDEOPHY, we curate a wide range of prompts that would be used to assess the physical commonsense of the T2V models.

To evaluate the quality of a T2V generative model, Fréchet video distance (FVD) is traditionally used to measure the similarity between real and generated video distributions [77, 18]. However, FVD has several limitations for assessing physical commonsense including the requirement for a reference video that is difficult to obtain for novel scenes, bias towards video quality, and failure to detect unrealistic motions [15, 72]. Similarly, CLIPScore [61] measures *semantic* similarity between generated video frames and the conditioning text in a shared representation space, making it unsuitable for evaluating physical commonsense in generated videos. Moreover, prior work [32] introduced a comprehensive benchmark to evaluate various qualities of generated videos (e.g., motion smoothness, background consistency) using existing models, but it does not specifically address the generated videos' adherence to physical laws. Therefore, existing benchmarks and metrics are either unreliable or lack coverage for holistic evaluation of the physical commonsense capabilities.

To this end, we propose VIDEOPHY, a dataset designed to evaluate the adherence of generated videos to physical commonsense in real-world activities. Specifically, physical commonsense focuses on the intuitive understanding of the behavior and dynamics of various states of matter (solids, fluids) in the physical world [58, 91, 10]. For instance, 'water pouring into a glass' will intuitively result in the water level in the glass rising over time. As a result, we rely on human perception and experience in the physical world to assess the adherence of the generated videos to physical laws instead of precise dynamical equations, which are harder to assess. In Figure 2, we provide qualitative examples to illustrate physical commonsense violations in the videos. Our dataset is constructed through a three-stage pipeline that involves (a) prompting a large language model [53] to generate candidate captions that depict interactions between diverse states of matter (e.g., solid-solid, solid-fluid, fluid-fluid), (b) human verification of the generated captions, and (c) annotating the complexity in rendering objects or synthesizing motions described in the captions based on physics simulation.

In total, VIDEOPHY comprises 688 high-quality, human-verified captions that will be used to generate videos from T2V models. In addition, the dataset consists human-labeled annotations for physical commonsense of the generated videos. Specifically, we acquire generated videos from **nine** diverse T2V models including open models (e.g., OpenSora [55], StableVideoDiffusion [11], VideoCrafter2 [20]) and closed models (e.g., Pika [57], Lumiere [7] from Google, Gen-2 [24] from Runway). Subsequently, we perform human evaluation on the generated videos for semantic adherence to the conditioning text (e.g., do the videos follow the caption?) and physical commonsense (e.g., do the videos follow physical laws intuitively?). Interestingly, we find that the existing T2V generative models severely lack the capability to follow caption accurately and generate videos with physical commonsense. Specifically, the best performing model, Pika, follows the text and generates physically accurate videos for 19.7% of the instances (§5). In Figure 1, we compare the performance (i.e., accurate semantic adherence and physical commonsense) of various T2V generative models on the VIDEOPHY dataset, as judged by human annotators.

Although human evaluation of semantic adherence and physical commonsense is reliable, it is both expensive and difficult to scale. To address this challenge, we introduce VIDEOCON-PHYSICS, a video-language model designed to assess the semantic adherence and physical commonsense of generated videos using user queries grounded in text. Specifically, we fine-tune VIDEOCON [3], a robust semantic adherence evaluator for real videos, on generated videos and human annotations from our VIDEOPHY dataset. Our results demonstrate that VIDEOCON-PHYSICS outperforms Gemini-Pro-Vision-1.5 [63], showing a 9 points improvement in semantic adherence and a 15 points improvement in physical commonsense on unseen prompts. Overall, the VIDEOPHY dataset aims to bridge the gap in understanding physical commonsense in generated videos and enables scalable testing.

2 VIDEOPHY Dataset

Our dataset, VIDEOPHY, aims to offer a robust evaluation benchmark for physical commonsense in video generative models. Specifically, the dataset is curated with guidelines to cover (a) a wide range of daily activities and objects in the physical world (e.g., rolling objects, pouring liquid into a glass), (b) physical interactions between various material types (e.g., solid-solid or solid-fluid interactions), and (c) the perceived complexity of rendering objects and motions under graphic simulation. For instance, *ketchup*, which follows Non-Newtonian fluid dynamics [85], is harder to model and simulate than *water*, which follows Newtonian fluid dynamics, using traditional fluid simulators [14].

Under the collection guidelines, we curate a list of text prompts that will be used for conditioning the text-to-video generative models. Specifically, we follow the 3-stage pipeline to create the dataset.

LLM-Generated Captions (Stage 1). Here, we query a large language model, in our case GPT-4 [53], to generate a list of 1000 *candidate* captions depicting real-world dynamics. As the majority of real-world dynamics involve solids or fluids, we broadly classify those dynamics into three categories: *solid-solid* interactions, *solid-fluid* interactions, and *fluid-fluid* interactions. Specifically, we consider fluid dynamics involving in-viscid and viscous flows—representative examples being water and honey, respectively. On the other hand, we find that solids exhibit more diverse constitutive models, including but not limited to rigid bodies, elastic materials, sands, metals, and snow. In total, we prompt GPT-4 to generate 500 candidate captions for solid-solid and solid-fluid interactions, and 200 candidate captions for fluid-fluid interactions. We present the GPT-4 prompts in Appendix G.

Human Verification (Stage 2). Since LLM-generated captions may not adhere to our input query, we perform a human verification step to filter bad generations. Specifically, the authors perform human verification to ensure the quality and relevance of the captions, adhering to these criteria: (1) the caption must be clear and understandable (2) the caption should avoid excessive complexity, such as overly varied objects or too intricate dynamics (3) the captions must accurately reflect the intended interaction categories, ensuring, for example, that fluids are indeed described in solid-fluid or fluid-fluid dynamics. To maintain focus on the fundamental interactions among solids and fluids, we also exclude captions involving complex physical phenomena such as phase changes (e.g. ice melting into water) or magnetic effects. Finally, we have 688 captions where 289 captions for solid-solid interactions, 291 for solid-fluid interactions, and 108 for fluid-fluid interactions, respectively.

Difficulty Annotation (Stage 3). To acquire fine-grained insights into the quality of the video generation, we further annotate our each instance in the dataset with perceived *difficulty*. Specifically, we ask two experienced graphics researchers (senior Ph.D. students in physics-based simulation) to independently classify each caption as easy (0) or hard (1) based on their perception of the complexity in simulating the objects and motions in the captions using state-of-the-art physics engines [41, 22, 86, 96, 60, 26]. Subsequently, the disagreements were discussed to reach a unanimous judgement for less than 5% of the instances. We note that the level of difficulty is evaluated within each category (e.g., solid-solid, solid-fluid, fluid-fluid), and cannot be compared across different categories. We present the examples for generated captions in Table 4 in Appendix C.

Data Analysis. A fine-grained metadata facilitates a comprehensive understanding of the benchmark. Specifically, we present the main statistics of the VIDEOPHY dataset in Table 5 in Appendix E. Notably, we generate 9000+ videos for the prompts in the dataset using a diverse range of generative models. In addition, the average caption length is 8.5 words, indicating that most captions are straightforward and do not complicate our analysis with complex phrasing that could be excessively challenging the generative models. The dataset includes 138 unique actions grounded in our captions. Additionally, Appendix Figure 4 visualizes the root verbs and direct nouns used in the VIDEOPHY captions, highlighting the diversity of actions and entities depicted. Hence, our dataset encompasses a wide range of visual concepts and actions. We present a fine-grained data analysis in Appendix J.

3 Evaluation

3.1 Metrics

The ability to assess the quality of the generated videos is a challenging task. While humans can evaluate videos across various visual dimensions [32, 18], we focus primarily on the models' adherence to the provided text and the incorporation of physical commonsense. These are key objectives that conditional generative models must maximize.

Semantic Adherence (SA). This metric evaluates whether the text caption is semantically grounded in the frames of the generated videos. Specifically, it assesses if the actions, events, entities, and their relationships are perceived to be correctly depicted in the video frames (e.g., water is flowing into the glass in the generated video for the caption 'water pouring into the glass'). In this work, we annotate the generated videos for semantic adherence, denoted as SA $= \{0, 1\}$. Here, SA $= 0$ indicates that some or all of the caption is not grounded in the generated video.

Physical Commonsense (PC). This metric evaluates whether the depicted actions, and object's state follow the physics laws in the real-world. For instance, the level of water should increase in the glass as water flows into it, following conversation of mass. In this work, we annotate the physical commonsense of the generated videos, denoted as PC $= \{0, 1\}$. Here, PC $= 1$ indicates that the generated movements and interactions align with intuitive physics that humans acquire with their experience in the real-world. As physical commonsense is entirely grounded in the video, it is independent of the semantic adherence capability of the generated video. In this work, we compute the fraction of the videos for which semantic adherence is high (SA $= 1$), physical commonsense is high (PC $= 1$), and joint performance of these metrics is high (SA $= 1,$ PC $= 1$).

3.2 Human Evaluation

We conducted a human evaluation to assess the performance of the generated videos in terms of semantic adherence and physical commonsense using our dataset. Annotations were obtained from a group of qualified Amazon Mechanical Turk (AMT) workers who had passed a qualification test. The workers were compensated at a rate of $\$18$ per hour. In this task, annotators were presented with a caption and the corresponding generated video without any information about the generative model. They were asked to provide a semantic adherence score (0 or 1) and a physical commonsense score (0 or 1) for each instance. Annotators were instructed to treat semantic adherence and physical commonsense as independent metrics and were shown several solved examples by the authors before starting the main annotation task. In some cases, we find that generative models create static scenes instead of video frames with high motion. Here, we ask annotators to judge the physical plausibility of the static scene in the real world (e.g., a static scene of a folded brick does not follow physical

commonsense). However, if the static scenes are noisy (e.g., unwanted grainy or speckled patterns), we instruct them to consider it as poor physical commonsense.

In our experiments, the annotators have studied high school-level physics. However, the human annotators were not asked to list the violation of the physics laws since it would make the annotations more time-consuming and expensive. Additionally, the current annotations can be performed by annotators experience in the physical world (e.g., workers know that water flows *down* from a tap, shape of a wood log *will not change* while floating on water) instead of advanced education in physics. A screenshot of the human annotation interface is presented in Appendix H.

3.3 Automatic Evaluation

While the human evaluation is more accurate for model benchmarking, it is time-consuming and expensive to collect at scale. To this end, we evaluate the performance of various zero-shot methods in judging the quality of the generated videos in terms of semantic adherence and physical commonsense. Further, we propose VIDEOCON-PHYSICS, a capable automatic evaluator on our dataset.

Baselines. Similar to [4], we utilize the capability of **GPT-4Vision** [54] to reason over multiple images in a zero-shot manner. Specifically, we prompt the GPT-4V model with the caption and 8 video frames sampled uniformly from the generated video. Here, we instruct the model to provide the semantic adherence (0 or 1) and physical commonsense score (0 or 1). Since GPT-4V does not process videos natively, we assess the automatic evaluation using **Gemini-Pro-Vision-1.5**, which can input the caption and the entire generated video. Specifically, we instruct it to provide the semantic adherence (0 or 1) and physical commonsense (0 or 1) of the input video, identical to the GPT-4V analysis. We provide the prompts used in the experiments in Appendix I.

Since the previous two models are closed, it is difficult to fine-tune them with custom data. As a result, we use **VIDEOCON**, an open generative video-text language model with $7B$ parameters, that is trained on real videos for robust semantic adherence evaluation [3]. Specifically, we prompt VIDEOCON to generate a text response (*Yes*/*No*) conditioned on the multimodal template $\mathcal{T}_t(x)$ for semantic adherence and physical commonsense tasks. Formally,

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

where t is either semantic adherence to the caption or physical commonsense task, C is the conditioning caption and V is the generated video for the caption C. We provide the multimodal templates ($\mathcal{T}_{SA}(V, C)$, $\mathcal{T}_{PC}(V)$) in Appendix I. We compute the score from the VIDEOCON model p_θ:

$$s_\theta(\mathcal{T}_t(x)) = \frac{p_\theta(\textit{Yes}|\mathcal{T}_t(x))}{p_\theta(\textit{Yes}|\mathcal{T}_t(x)) + p_\theta(\textit{No}|\mathcal{T}_t(x))}, \tag{2}$$

where $p_\theta(\textit{Yes}|\mathcal{T}_t(x))$ is the probability of '*Yes*' conditioned on $\mathcal{T}_t(x)$, and $t \in \{SA, PC\}$. [1]

VIDEOCON-PHYSICS. Since VIDEOCON is not trained on the generated video distribution or equipped to judge physical commonsense, it is not expected to perform well in our setup in a zero-shot manner. To this end, we propose VIDEOCON-PHYSICS, an open-source generative video-text model, that can assess the semantic adherence and physical commonsense of the generated videos. Specifically, we finetune VIDEOCON by combining the human annotations acquired for the semantic adherence and physical commonsense tasks over the generated videos.[2]

Overall, we evaluate the usefulness of the baseline and the proposed model by computing the AUC-ROC between the human annotations and model predictions for diverse generated videos on the unseen prompts. We will provide more details on the train and test set in §4.2.

[1] As a large video multimodal model, VIDEOCON predicts a token distribution over the entire token vocabulary conditioned on the multimodal template. Therefore, $p_\theta(\textit{Yes}|\mathcal{T}_t(x)) + p_\theta(\textit{No}|\mathcal{T}_t(x))$ is not equal to 1.

[2] Finetuning a separate classifier for semantic adherence and physical commonsense did not provide any additional benefits over a single classifier (VIDEOCON-PHYSICS) trained in a multi-task manner.

4 Setup

In this section, we present the list of text-to-video generative models benchmarked on the VIDEOPHY dataset (§4.1) and provide further details about the dataset splits (§4.2).

4.1 Text-to-Video Generative Models

We evaluate a diverse range of **nine** closed and open text-to-video generative models on VIDEOPHY dataset. The list of the models includes *ZeroScope* [19], *LaVIE* [83], *VideoCrafter2* [20], *OpenSora* [55], *StableVideoDiffusion (SVD)-T2I2V* [11], *Gen-2 (Runway)* [24], *Lumiere-T2V*, *Lumiere-T2I2V* (Google) [7], and *Pika* [57]. Here, the *T2I2V* models involve the generation of an image (*I*) conditioned on the caption (*T*) followed by video generation (*V*) conditioned on the generated image. We provide more details about these models in Appendix F. While there are various closed models such as Sora [46] and Genmo [28], we could not get access through their videos due to the lack of API support. We provide inference details in Appendix L.

4.2 Dataset

As described earlier, we train an automatic evaluation model to enable cheaper and scalable testing of the generated videos on our dataset (§ 3.3). To facilitate this, we split the prompts in the VIDEOPHY dataset equally into *train* and *test* sets. Specifically, we utilize the human annotations on the generated videos for the 344 prompts in the *test* set for benchmarking, while the human annotations on the generated videos for the 344 prompts in the *train* set are used for training the automatic evaluation model. We ensure that the distribution of the state of matter (solid-solid, solid-fluid, fluid-fluid) and complexity of the captions (easy, hard) is similar in the training and testing.

Benchmarking. Here, we generate one video per test prompt for each T2V generative model in our testbed. Subsequently, we ask three human annotators to judge the semantic adherence and physical commonsense of the generated videos. In our experiments, we report the majority-voted scores from the human annotators. We find that the inter-annotator agreement for semantic adherence and physical commonsense judgment is 75% and 70%, respectively. This indicates that the human annotators find the task of judging physical commonsense more subjective than semantic adherence. [3] In total, we collect $18500+$ human annotations across the testing prompts and T2V models.

Training set for VIDEOCON-PHYSICS. Here, we sample two videos per training prompt for each T2V generative model in our testbed. Specifically, we choose two videos to obtain more data instances for training the automatic evaluation model. Subsequently, we ask one human annotator to judge the semantic adherence and physical commonsense of the generated videos. In total, we collect $12000+$ human annotations, half of them for semantic adherence and the other half for physical commonsense. Specifically, we finetune VIDEOCON to maximize the log likelihood of *Yes*/*No* conditioned on the multimodal template for semantic adherence and physical commonsense tasks (Appendix I). We do not collect three annotations per video as it is financially expensive. In total, we spent $\$2800$ on collecting human annotations for benchmarking and training.

5 Results

Here, we present the results of the T2V generative models (§5.1), and establish the effectiveness of the VIDEOCON-PHYSICS as an automatic evaluator on the VIDEOPHY dataset (§3.3).

5.1 Performance on VIDEOPHY Dataset

We compare the performance of the T2V generative models on the VIDEOPHY dataset using human evaluation in Table 1. Specifically, we find that the Pika (closed model) and VideoCrafter2 (open model) generates videos that adhere to the caption and follow physics laws (SA $= 1$, PC $= 1$) in

[3] Since most of the generated videos are not perfect, the variations in the annotations result from diverse tolerance for physical laws violations. As the generative models improve, we believe that the human annotations will achieve higher agreement on our dataset.

Table 1: **Human evaluation results on the VIDEOPHY dataset.** We report the percentage of testing prompts for which the T2V models generate videos that adhere to the conditioning caption and exhibit physical commonsense. We abbreviate semantic adherence as SA, and physical commonsense as PC. **SA, PC** indicates the percentage of the instances for which SA=1 and PC=1. Ideally, we want the generative models to maximize the performance on this metric. In the first column, we highlight the overall performance, and the later columns are dedicated to fine-grained performance for the interaction between different states of matter in the prompts.

	Overall (%)			Solid-Solid (%)			Solid-Fluid (%)			Fluid-Fluid (%)		
Model	**SA, PC**	**SA**	**PC**	**SA, PC**	**SA**	**PC**	**SA, PC**	**SA**	**PC**	**SA, PC**	**SA**	**PC**
Open Models												
VideoCrafter2 [20]	19.0	48.5	34.6	4.9	31.5	23.8	27.4	57.5	41.8	32.7	69.1	43.6
LaVIE [83]	15.7	48.7	28.0	8.5	37.3	19.0	15.8	52.1	30.8	34.5	69.1	43.6
SVD-T2I2V [12]	11.9	42.4	30.8	4.2	25.9	27.3	17.1	52.7	32.9	18.2	58.2	34.5
ZeroScope [19]	11.9	30.2	32.6	6.3	17.5	22.4	14.4	40.4	37.0	20.0	36.4	47.3
OpenSora [55]	4.9	18.0	23.5	1.4	7.7	23.8	7.5	30.1	21.9	7.3	12.7	27.3
Closed Models												
Pika [57]	19.7	41.1	36.5	13.6	24.8	36.8	16.3	46.5	27.9	44.0	68.0	58.0
Lumiere-T2I2V [7]	12.5	48.5	25.0	8.4	37.1	25.2	17.1	59.6	26.0	10.9	49.1	21.8
Lumiere-T2V [7]	9.0	38.4	27.9	8.4	26.6	27.3	9.6	47.3	26.0	9.1	45.5	34.5
Gen-2 (Runway) [24]	7.6	26.6	27.2	4.0	8.9	37.1	8.1	38.5	18.5	15.1	37.7	26.4

19.7% and 19% of the cases, respectively. This indicates that the video generative models struggle on the VIDEOPHY dataset, and far from being general-purpose simulators of the physical world. Pika stands out as the best model for generating videos that demonstrate physical commonsense, achieving a performance of 36.5%, while VideoCrafter2 is a close second at 34.6%. VideoCrafter2's training process involves a mix of low-quality video-text data and high-quality image-text data, suggesting that incorporating high-quality video data could further enhance its performance. Amongst the closed models, Lumiere-T2I2V emerges as the top video generative model for producing videos that accurately follow the conditioning captions, with a performance rate of 56.6%. This indicates at the effectiveness of cascaded approach (text-to-image followed by image-to-video) in generating high-quality, text-adherent videos. Conversely, among the open models, OpenSora performs the worst on the VIDEOPHY dataset, indicating significant potential for the community to improve open-source implementations of Sora.

Table 2: **Fine-grained performance across caption complexity using human evaluation.** We find that T2V models struggle more on the harder captions than the easier captions in both the semantic adherence (SA) and physical commonsense (PC) metrics.

	Easy (%)		Hard (%)	
Model	**SA**	**PC**	**SA**	**PC**
Open Models				
VideoCrafter2	53.4	38.1	42.6	30.3
LaVIE	51.9	31.2	44.8	24.0
SVD-T2I2V	41.8	37.6	43.2	22.6
ZeroScope	32.3	33.9	27.7	31.0
OpenSora	20.1	25.4	5.2	21.3
Closed Model				
Pika	45.7	39.9	35.1	32.1
Lumiere-T2I2V	56.6	29.1	38.7	20.0
Lumiere-T2V	38.6	34.9	38.1	19.4
Gen-2 (Runway)	26.6	31.8	26.6	21.6

Variation with the states of matter. We study the variation in the performance of T2V models with the interaction between the diverse states of matter grounded in the captions (e.g., solid-solid) in Table 5.1. Interestingly, we find that all the existing T2V models perform the worst on the captions that depict interactions between solid materials (e.g., *bottle* topples off the *table*), with the best performing model, Pika, achieving 13.6% on accurate semantic adherence and physical commonsense. Furthermore, we observe that VideoCrafter2 achieves the highest performance in the captions that depict interaction between solid and fluid material types (e.g., a whisk mixes an egg). This indicates that the T2V model performance is greatly influenced by the states of matter involved in a scene, and highlights that model developers can focus on enhancing semantic adherence and physical commonsense for solid-solid interactions.

Variation with the complexity. We analyze the variation in the T2V model performance with the complexity in rendering objects or synthesizing interactions grounded in the caption under physical simulation in Table 2. We find that the semantic adherence and physical commonsense performance of all the T2V models decreases as the complexity of the captions increases. This indicates that the captions that are harder to simulate physically are also harder to control via conditioning for the T2V generative models. Our analysis thus highlights that the future T2V model development should focus

(a) v. SVD-T2I2V. *An apple bobbing in a bucket of water.*

(b) v. Lumiere-T2I2V. *Red wine swirling in water.*

(c) v. Gen-2. *Cymbal clashing with wooden drumstick.*

Figure 3: **Qualitative comparison of Pika with other models.** The top row presents the generated videos from the Pika model. (a) For SVD-T2I2V, the apple remains unnaturally still in the flowing water. (b) For Lumiere-T2I2V, a part of the red wine is not falling. (c) For Gen-2, the drumstick deforms over time.

on reducing the gap between the easy and the hard captions from our VIDEOPHY dataset. We present the results for additional metrics in Table 7 in Appendix K.

Table 3: **Comparison of ROC-AUC for automatic evaluation methods.** We find that VIDEOCON-PHYSICS outperforms diverse baselines, including GPT-4Vision and Gemini-1.5-Pro-Vision, for semantic adherence (SA) and physical commonsense (PC) judgments on the testing prompts.

Method	ROC-AUC SA	ROC-AUC PC
Random	50	50
GPT-4-Vision [54]	53	53
Gemini-Pro-Vision-1.5 [63]	73	58
VideoCon [3]	65	54
VIDEOCON-PHYSICS (Ours)	82	73

5.2 VIDEOCON-PHYSICS: Automatic Evaluator for VIDEOPHY Dataset

Here, we propose VIDEOCON-PHYSICS model for scalable and reliable evaluation of semantic adherence and physical commonsense in the generated videos. We compare the ROC-AUC agreement of different automatic evaluators with the human predictions on the testing prompts in Table 3. We find that the VIDEOCON-PHYSICS outperforms the zeroshot VIDEOCON by 17 points and 19 points on the semantic adherence and physical commonsense judgment, respectively. This highlights that finetuning with the generated video distribution and human annotations aids in improving the model judgment on the unseen prompts. Further, we notice that the model's agreement are higher for semantic adherence as compared to the physical commonsense. This indicates that judging physical commonsense is a harder task than judging semantic adherence for VIDEOCON-PHYSICS.

Interestingly, we observe that the GPT-4-Vision's judgments are close to random for semantic adherence and physical commonsense on our dataset. This implies that faithful evaluations are hard to obtain from the multi-image reasoning capabilities of the GPT-4-Vision in a zeroshot manner. To address this, we test Gemini-Pro-Vision-1.5 and find that it achieves a good semantic adherence score (73 points), however, it is close to random in physical commonsense evaluation (54 points). This highlights that the existing multimodal foundation models lack the capability to judge physical commonsense. In Appendix N, we show that VIDEOCON-PHYSICS also generalizes to unseen generative models, suggesting that is the best automatic evaluator for the VIDEOPHY dataset. We provide more discussion on the usefulness of VIDEOCON-PHYSICS in Appendix O.

6 Qualitative Examples

Here, we present some qualitative examples to understand the common failure modes in the generated video regarding poor physical commonsense. Qualitative examples from various T2V generative models are provided in Figure 15 - 23 in Appendix Q. The common failure modes include – (a) *Conservation of mass violation*: the volume or texture of an object is not consistent over time, (b) *Newton's First Law violation*: an object changes its velocity in a balanced state without any external

force, (c) *Newton's Second Law violation*: an object violates the conversation of momentum, (d) *Solid Constitutive Law violation*: solids deform in ways that contradict their material properties, e.g., a rigid object deforming over time, (e) *Fluid Constitutive Law violation*: fluids exhibit unnatural flow motions, and (f) *Non-physical penetration*: objects unnaturally penetrate each other.

In addition, we analyze some qualitative examples to understand the gap between the top-tier models (Pika and VideoCrafter2) and the other models in our testbed. We present the examples in Figure 3 and Figure 14 in Appendix P. For instance, we find that SVD-T2I2V is likely to underperform in scenes involving vibrant fluid dynamics. Lumiere-T2I2V performs better than Lumiere-T2V in terms of visual quality, but still lacks a profound understanding of gravity (e.g. in Figure 3(b)). Gen-2 sometimes cannot differentiate multiple objects, thus deforming rigid objects in dynamic motions. Additional observations are reported in Appendix P. Our analysis highlights the lack of fine-grained physical commonsense understanding that future video modeling research should aim to address.

7 Related Work

Video Generation Models. Recent advancements in video generation models have emerged from two primary architectures: diffusion-based models [24, 12, 46, 13, 83, 81, 20, 34] and autoregressive modeling-based approaches [92, 37, 30, 78]. Among these, diffusion models have garnered significant attention. The model known as SVD [12], built on a latent diffusion model (LDM) [64], proposes a three-stage training process for video LDMs. Given the rapid development of video generation technology, our work focuses on evaluating video generation models for their physical commonsense.

Evaluating Video Generation Models. Traditional evaluation methods for video generation primarily employ metrics such as FVD [77] and IS [67]. However, there is a growing consensus on the need for more comprehensive metrics to assess the performance of video generation models [32, 45, 39, 44]. V-Bench [32] offers a detailed benchmark suite that introduces a hierarchical evaluation protocol with granular perspectives. Here, we propose VIDEOPHY dataset to measure the physical commonsense in the generated videos. Additionally, we introduce a VIDEOCON-PHYSICS auto-evaluator and analyze specific physical laws [10] that are violated in the generated videos through qualitative analysis. We present a detailed related work in Appendix D.

8 Conclusion

In this work, we introduce VIDEOPHY, a first of its kind dataset to assess the physical commonsense in the generated videos. Further, we evaluate a diverse set of T2V models (open and closed models) and found that they significantly lack in the physical commonsense and semantic adherence capabilities. Our dataset unveils that the existing methods are far being general-purpose world simulators. Further, we introduce VIDEOCON-PHYSICS, an auto-evaluation model that enables cheap and scalable evaluation on our dataset. We believe that our work will serve as the cornerstone in studying physical commonsense for video generative modeling.

9 Acknowledgement

Hritik Bansal is supported in part by AFOSR MURI grant FA9550-22-1-0380.

References

[1] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video generative adversarial networks: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–25, 2022.

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[3] Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. *arXiv preprint arXiv:2311.10111*, 2023.

[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.

[5] Hritik Bansal, Ashima Suvarna, Gantavya Bhatt, Nanyun Peng, Kai-Wei Chang, and Aditya Grover. Comparing bad apples to good oranges: Aligning large language models via joint preference optimization. *arXiv preprint arXiv:2404.00530*, 2024.

[6] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022.

[7] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[8] David Baraff. An introduction to physically based modeling: rigid body simulation i—unconstrained rigid body dynamics. *SIGGRAPH course notes*, 82, 1997.

[9] Christopher Batty, Florence Bertails, and Robert Bridson. A fast variational framework for accurate solid-fluid coupling. *ACM Transactions on Graphics (TOG)*, 26(3):100–es, 2007.

[10] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[11] 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.

[12] 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.

[13] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[14] Robert Bridson. *Fluid simulation for computer graphics*. AK Peters/CRC Press, 2015.

[15] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022.

[16] 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.

[17] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

[18] Emanuele Bugliarello, H Hernan Moraldo, Ruben Villegas, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Han Zhang, Dumitru Erhan, Vittorio Ferrari, Pieter-Jan Kindermans, and Paul Voigtlaender. Storybench: A multifaceted benchmark for continuous story visualization. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] cerspense. cerspense/zeroscope_v2_576w · Hugging Face — huggingface.co. `https://huggingface.co/cerspense/zeroscope_v2_576w`, 2023.

[20] 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. *arXiv preprint arXiv:2401.09047*, 2024.

[21] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024.

[22] Yunuo Chen, Tianyi Xie, Cem Yuksel, Danny Kaufman, Yin Yang, Chenfanfu Jiang, and Minchen Li. Multi-layer thick shells. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023.

[23] 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.

[24] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[25] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[26] Yu Fang, Ziyin Qu, Minchen Li, Xinxin Zhang, Yixin Zhu, Mridul Aanjaneya, and Chenfanfu Jiang. Iq-mpm: an interface quadrature material point method for non-sticky strongly two-way coupled nonlinear solids and fluids. *ACM Transactions on Graphics (TOG)*, 39(4):51–1, 2020.

[27] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[28] genmo. Genmo. Create videos and images with AI. — genmo.ai. `https://www.genmo.ai/`.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[30] 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.

[31] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[32] 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. *arXiv preprint arXiv:2311.17982*, 2023.

[33] huggingfaceEulerDiscreteScheduler. EulerDiscreteScheduler — huggingface.co. `https://huggingface.co/docs/diffusers/en/api/schedulers/euler`.

[34] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

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

[36] 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.

[37] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[38] Dan Koschier, Jan Bender, Barbara Solenthaler, and Matthias Teschner. Smoothed particle hydrodynamics techniques for the physics based simulation of fluids and solids. *arXiv preprint arXiv:2009.06944*, 2020.

[39] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dateset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024.

[40] Egor Larionov, Christopher Batty, and Robert Bridson. Variational stokes: a unified pressure-viscosity solver for accurate viscous liquids. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

[41] Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy R Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M Kaufman. Incremental potential contact: intersection- and inversion-free, large-deformation dynamics. *ACM Trans. Graph.*, 39(4):49, 2020.

[42] Minchen Li, Danny M Kaufman, and Chenfanfu Jiang. Codimensional incremental potential contact. *arXiv preprint arXiv:2012.04457*, 2020.

[43] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465*, 2024.

[44] 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. *arXiv preprint arXiv:2404.01291*, 2024.

[45] 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. *arXiv preprint arXiv:2310.11440*, 2023.

[46] 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.

[47] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

[48] Mariem Mezghanni, Malika Boulkenafed, Andre Lieutier, and Maks Ovsjanikov. Physically-aware generative network for 3d shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9330–9341, 2021.

[49] mplugowl. mplug-owl-video. `https://github.com/X-PLUG/mPLUG-Owl/tree/main/mPLUG-Owl/mplug_owl_video`.

[50] Matthias Müller, Barbara Solenthaler, Richard Keiser, and Markus Gross. Particle-based fluid-fluid interaction. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 237–244, 2005.

[51] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[52] 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.

[53] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774, 2023a*, 2023.

[54] OpenAI. Gpt-4v(ision) system card, 2023b. `https://openai.com/research/gpt-4v-system-card`, 2023.

[55] OpenSora. GitHub - hpcaitech/Open-Sora: Open-Sora: Democratizing Efficient Video Production for All — github.com. `https://github.com/hpcaitech/Open-Sora`, 2024.

[56] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[57] pika. Pika — pika.art. `https://pika.art/`.

[58] Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9):1257–1267, 2022.

[59] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[60] Ziyin Qu, Minchen Li, Yin Yang, Chenfanfu Jiang, and Fernando De Goes. Power plastics: A hybrid lagrangian/eulerian solver for mesoscale inelastic flows. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[61] 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.

[62] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[63] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[66] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[67] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[68] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[69] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020.

[70] Eftychios Sifakis and Jernej Barbic. Fem simulation of 3d deformable solids: a practitioner's guide to theory, discretization and model reduction. In *Acm siggraph 2012 courses*, pages 1–50. 2012.

[71] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[72] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022.

[73] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[74] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[75] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[76] Tetsuya Takahashi, Tomoyuki Nishita, and Issei Fujishiro. Fast simulation of viscous fluids with elasticity and thermal conductivity using position-based dynamics. *Computers & Graphics*, 43:21–30, 2014.

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

[78] 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 descriptions. In *International Conference on Learning Representations*, 2022.

[79] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

[80] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024.

[81] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[82] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023.

[83] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[84] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[85] Mariusz Witczak, Lesław Juszczak, and Dorota Gałkowska. Non-newtonian behaviour of heather honey. *Journal of Food Engineering*, 104(4):532–537, 2011.

[86] Tianyi Xie, Minchen Li, Yin Yang, and Chenfanfu Jiang. A contact proxy splitting method for lagrangian solid-fluid coupling. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.

[87] Tianyi Xie, Zeshun Zong, Yuxin Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023.

[88] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023.

[89] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022.

[90] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[91] Ilker Yildirim, Max H Siegel, Amir A Soltani, Shraman Ray Chaudhuri, and Joshua B Tenenbaum. Perception of 3d shape integrates intuitive physics and analysis-by-synthesis. *Nature Human Behaviour*, 8(2):320–335, 2024.

[92] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[93] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023.

[94] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. *arXiv preprint arXiv:2404.13026*, 2024.

[95] Jing Zhou, Zongyu Lin, Yanan Zheng, Jian Li, and Zhilin Yang. Not all tasks are born equal: Understanding zero-shot generalization. In *The Eleventh International Conference on Learning Representations*, 2022.

[96] Zeshun Zong, Xuan Li, Minchen Li, Maurizio M Chiaramonte, Wojciech Matusik, Eitan Grinspun, Kevin Carlberg, Chenfanfu Jiang, and Peter Yichen Chen. Neural stress fields for reduced-order elastoplasticity and fracture. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

A Limitations

In this work, we evaluate the physical commonsense capabilities of T2V generative models. Specifically, we curated the VIDEOPHY dataset, consisting of 688 captions. We argue that the captions are comprehensive and high-quality after going through our three-stage data curation pipeline. In the future, it will be pertinent to expand the physical commonsense understanding to more branches of physics, including projective geometry. Additionally, we test a diverse set of T2V generative models, including both open and closed models. While it is financially and computationally challenging to evaluate an exhaustive list of models, we have aimed to incorporate models with diverse architectures, training datasets, and inference strategies. In the future, it will be important to gain access to and include new high-performance T2V models in our study.

In addition, we perform human annotations using Amazon Mechanical Turkers (AMT), where most of the workers primarily belong to the US and Canada. Hence, the human annotations in this work do not represent the diverse demographics around the globe. As a result, our human annotations reflect the perceptual biases of the annotators from Western cultures. In the future, it will be pertinent to assess the impact of diverse groups on our human evaluations. Finally, we acknowledge that text-to-video generative models can perpetrate societal biases in their generated content [80, 6]. It is critical that future work quantifies this bias in the generated videos and provides methods for the safe deployment of the models.

B Data Licensing

The VIDEOPHY dataset comprises videos generated by various T2V (Text-to-Video) generative models, detailed in Section F. The licensing terms for these videos will align with those specified by the respective model owners, as cited in this work. The curated captions and human annotations will be licensed under the MIT License.

C Example captions in the VIDEOPHY dataset

We present example captions in our dataset in Table 4.

Category	Difficulty	Example Captions
Solid-Solid	Easy	1. Bottle topples off the table. (rigid bodies)
		2. Ball bounces off the floor. (deformable and rigid bodies)
	Hard	1. Scrubber scrubs a dirty dish. (complex contacts)
		2. Scissors trim the paper. (material fracture)
Solid-Fluid	Easy	1. Water flows down a circular drain. (contacts with rigid bodies)
		2. A paint roller coating a wall. (slow/static fluids)
	Hard	1. A swimmer splashing in the sea water. (contacts with high-speed)
		2. A whisk mixes an egg in a bowl. (contacts with high-speed)
Fluid-Fluid	Easy	1. Rain splashing on a pond. (mixing of same fluids)
		2. Sour cream swirls in hot soup. (layering)
	Hard	1. Ink spreading in still water. (mixing of different fluids)
		2. Honey threading slowly into tea. (mixing of different fluids)

Table 4: **Example captions in the VIDEOPHY dataset.** Specifically, we design them to depict the interactions between two states of matter (solid-solid, solid-fluid, fluid-fluid). We further classify the captions as easy or hard based on the modeling and simulation complexity in the computer graphics. We highlight the reasoning behind the easy and hard annotations by our annotators in the yellow.

D Detailed Related Work

Video Generation Models. Recent advancements in video generation models have emerged from two primary architectures: diffusion-based models [24, 12, 46, 13, 83, 81, 20, 34] and autoregressive modeling-based approaches [92, 37, 30, 78]. Among these, diffusion models have garnered significant attention. The model known as SVD [12], built on a Latent Diffusion Model (LDM) [64], proposes a three-stage training process for video LDMs: text-to-image pretraining, video pretraining, and video finetuning. Sora [46] represents a state-of-the-art in video generation, utilizing a diffusion-transformer architecture with unified training recipes and enhancements in language description processing for video generation. ModelScope [81] is also a diffusion-based text-to-video model which combines a VQGAN [25], a text-encoder, and a denoising UNet. Another diffusion model, VideoCrafter2 [20], leverages low-quality videos and high-quality videos to generate high-quality videos. LaVIE [83] is composed of a base text-to-video model, a temporal interpolation model, and a video super-resolution model, indicating that joint image and video training and temporal self-attention with rotary positional embeddings are key components to boost performance. Given the rapid development of video generation technology, an effective evaluation method for the generated videos becomes crucial. Our paper focuses on evaluating text-to-video generation models for their physical commonsense capabilities.

Evaluating Video Generation Models. Traditional evaluation methods for video generation primarily employ metrics such as FVD [77] and IS [67]. However, there is a growing consensus on the need for more comprehensive metrics to assess the performance of video generation models [32, 45, 39, 44]. V-Bench [32] offers a detailed benchmark suite that introduces a hierarchical evaluation protocol, breaking down 'video generation quality' into various granular perspectives. Another framework, EvalCrafter [45], proposes 17 objective metrics. Despite these advancements, existing methods largely overlook the fundamental aspect of physical commonsense. Unlike static images, videos incorporate a temporal dimension, embedding physical commonsense information across frames. Our research dives into the measurement of physical commonsense [10] in videos. Additionally, we introduce a VIDEOCON-PHYSICS auto-evaluator and analyze specific physical laws that are violated in the generated videos through qualitative analysis.

Physics Modeling. Simulating physical behaviors of solids and fluids has always been an important and popular topic in computer graphics. For solid materials, the simplest physical model is the long-established rigid body simulation [8], where solids are assumed not to deform. Simulation of deformable solids [70], on the other hand, takes into account the strain and stress during deformation. To capture more complicated materials, researchers have been proposing increasingly intricate models for different materials, such as metal [52], sand [36], and snow [74]. In contrast, most of the common fluids [14] in daily life can be broadly categorized as inviscid [38], e.g., water and air, and viscous fluids [76, 40], e.g., honey and oil. Additionally, an orthogonal research direction is to accurately, efficiently, and robustly model contact and interaction between different materials. These include solid-solid [41, 42], solid-fluid [9, 86], and fluid-fluid interactions [50]. Further, recent advancements in computer vision have started exploring incorporating physics priors into various 3D-aware generation tasks to enhance physical plausibility, such as human animation [93, 69, 88] and 3D/4D generation [48, 87, 94]. In this work, instead of generating, we focus on identifying whether the generated video adheres to physical laws.

E Data Analysis

We present the data analysis in Table 5 and Figure 4.

F Video Generative Models

For the open models, we benchmark *Zeroscope* [19, 81], a latent diffusion-based text-to-video model that adapts the text-to-image generative model [65] for video generation by training on high-quality video and image data for enhanced visual quality. Further, we benchmark *LaVIE* [83], a cascaded video latent diffusion model instead of a single diffusion model. Specifically, the LaVIE model is trained with a specialized curated dataset for enhanced visual quality and diversity. In addition, we test *VideoCrafter2*, a latent diffusion T2V model that enhances video generation quality by

Table 5: Key statistics of the VIDEOPHY dataset.

Statistic	Number
Total captions	688
Unique actions	138
Total T2V models	9
Total generated videos	9300
Human annotations	30500
Category (Interacting materials)	3
Solid-Solid	289
Solid-Fluid	291
Fluid-Fluid	108
Category (Interaction complexity)	2
Easy	366
Hard	322
Average length of caption	8.5

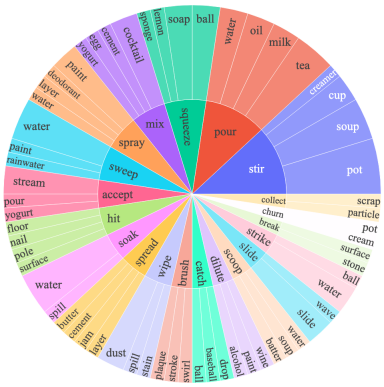


Figure 4: Top 20 most frequently occurring verbs (inner circle) and their top 4 direct nouns (outer circle) in our collected captions.

training on high-quality image-text data [75]. In our study, we also benchmark *OpenSora* [55], an open-source effort to replicate Sora [16], a high-performant closed latent diffusion model that uses diffusion transformers [56] for text-to-video generation. Finally, we include *StableVideoDiffusion (SVD)* [11], a latent diffusion model that can generate high resolution videos conditioned on a text or image. Since SVD-I2V (Image-to-Video) is publicly available, we utilize that to generate the videos. Specifically, we utilize SD-XL-Base-1.0 [59] to generate the conditioning images from the captions in the VIDEOPHY dataset. We term the entire pipeline as *SVD-T2I2V*.

For the closed models, we include *Gen-2* [24], a closed latent video diffusion model from Runway. In addition, we include Pika [57] with undisclosed information about the underlying generative model. Specifically, we wrote a custom API to acquire Gen-2 and Pika videos after paying for their monthly subscription for a total of $225. Finally, we include two versions of the *Lumiere* [7] from Google research. Specifically, *Lumiere-T2V* generates a video conditioned on the text, while *Lumiere-T2I2V* generates a video conditioned on an image, that is in-turn generated with the caption using a text-to-image generative model [66].

G Querying GPT-4 for Prompt Generation

In this section we discuss the prompt we utilized to generate all the prompts including three physical interaction categories: solid-solid, solid-fluid, fluid-fluid for video generation, which is displayed in Table 5, Table 6 and Table 7.

H Human Annotation Screenshot

We display the screenshot of our human annotation system in Figure 8

I Multimodal Template for Prompting Models

We present the prompts used for the GPT4V, Gemini-1.5-Pro-Vision, VideoCon baselines, and VIDEOCON-PHYSICS for semantic adherence evaluation in Figure 9 and physical commonsense alignment in Figure 10.

J Fine-Grained Statistics of Collections Across Different Physical Interaction Categories

In this section, we visualize the fine-grained statistics of collections across different physical interaction categories (Figure 11 - Figure 13).

Develop unique and imaginative captions, each briefly describing the interaction between two different solid materials in a realistic scene. Each caption should consist of 7-10 words and clearly indicate the solids involved in the action.

Guidelines:

1. Focus on common solids used in everyday scenarios, avoiding rare or seldom-used materials.

2. Exclude actions like 'celebrating', 'arguing', or 'laughing' that do not clearly involve physical interaction between materials.

3. Avoid generating static scenes (e.g., 'Lid covers pot to retain heat', 'Stack of paper sits on the desk').

4. Avoid adding participle phrases (e.g., 'sweetening it', 'a creamy swirl', 'fizzing energetically') in the caption.

5. The captions should focus on the actions that require contact forces, or friction forces. Do not focus on the actions that require penetration forces.

6. Format each caption as follows: 'action': ACTION, 'solid 1': SOLID, 'solid 2': SOLID, 'caption': CAPTION

Bad Examples Of Captions (Do Not Generate Such Captions):
A diamond scratching glass. ## Scratching action that requires penetration
A key scratches the surface of a wooden table. ## Scratching action that requires penetration

Good Examples Of Captions:
A brick presses down on a metal can.
A snowball falls to the ground and splits apart.
A small red elastic ball stuck to the wall.

Figure 5: GPT-4 Prompt to Generate Solid-Solid Captions.

Develop unique and imaginative captions, showcasing interaction between a solid material with a fluid material, for generating a video. After crafting the caption, list the entities that act as solid and fluid in the caption.

Guidelines:

1. Focus on common solids and fluids used in everyday scenarios, avoiding rare or seldom-used materials.

2. Exclude actions like 'celebrating', 'arguing', or 'laughing' that do not clearly involve physical interaction between materials.

3. Avoid actions that execute state change from solid to fluid or vice-versa.

4. Avoid generating static scenes (e.g., 'Lid covers pot to retain heat').

5. Avoid adding participle phrases (e.g., 'sweetening it', 'a creamy swirl', 'fizzing energetically') in the caption.

6. The captions should focus on the actions that require contact forces, or friction forces. Do not focus on the actions that require penetration forces.

7. Format each caption as follows: 'action': ACTION, 'solid': SOLID, 'fluid': FLUID, 'caption': CAPTION

Bad Examples Of Captions (Do Not Generate Such Captions):
Sugar dissolves in water. ## dissolving action will not be visible in video
Sulfuric acid corroding metal. ## corrosion will not be visible in video
Water boiling in a pot. ## boiling action will not be visible in video

Good Examples Of Captions:
A dam break releases a massive flood.
An iron rod falls into the water.
A metal spoon stirs the honey in a cup.

Figure 6: GPT-4 Prompt to Generate Solid-Fluid Captions.

Develop unique and imaginative captions, each briefly describing the interaction between two different fluid materials in a realistic scene. Each caption should consist of 7-10 words and clearly indicate the fluids involved in the action.

Guidelines:

1. Focus on common fluids used in everyday scenarios, avoiding rare or seldom-used materials.

2. Exclude actions like 'celebrating', 'arguing', or 'laughing' that do not clearly involve physical interaction between materials.

3. Avoid generating static scenes (e.g., 'Lid covers pot to retain heat').

4. Avoid adding participle phrases (e.g., 'sweetening it', 'a creamy swirl', 'fizzing energetically') in the caption.

5. The captions should focus on the actions that require mixing and laying for liquid-liquid interactions, or some contact forces between liquid and gas.

6. Format each caption as follows: 'action': ACTION, 'fluid 1': FLUID, 'fluid 2': FLUID, 'caption': CAPTION

Bad Examples Of Captions (Do Not Generate Such Captions):
Juice solidifies around water in ice trays. ## solidification won't be visible in the video
Sugar disappears into stirring water. ## dissolving won't be visible in the video An acid and a base react to neutralize each other, forming water. ## chemical reactions are not visible in the video

Good Examples Of Captions:
The wind creating ripples across the surface of the lake.
Milk falls into a transparent cup of water.
Oil falls into a transparent cup of water.

Figure 7: GPT-4 Prompt to Generate Fluid-Fluid Captions.

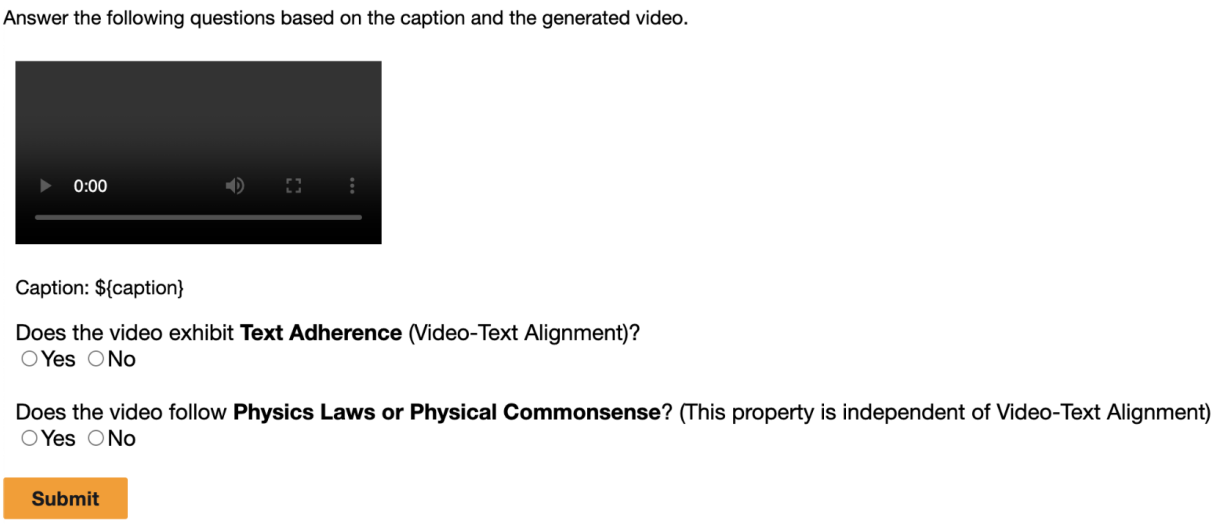


Figure 8: The screenshot of the human annotation interface.

K Fine-Grained Results

In this section, we report the fine-grained performance of semantic adherence and physical commonsense scores from all video generation models and compute the scores across different physical interaction categories (solid-solid, solid-fluid and fluid-fluid), as well as difficulty levels (0 and 1).

L Inference Details

We add the inference configurations for different video generation models in Table 8.

Table 6: Fine-grained performance of T2V models for the interaction between diverse states of matter using human evaluation. Ideally, we want the T2V models to achieve a high score on the *SA = 1 and PC = 1* metric while reduce the score on the *SA=0 and PC=0*, *SA=1 and PC=0*, and *SA=0 and PC=1* metrics.

Source	Category	SA (%)	PC (%)	SA=1 and PC=1 (%)	SA=1 and PC=0 (%)	SA=0 and PC=1 (%)	SA=0 and PC=0 (%)
Open Models							
LaVIE	Fluid-Fluid	69.1	43.6	34.5	34.5	9.1	21.8
	Solid-Fluid	52.1	30.8	15.8	36.3	15.1	32.9
	Solid-Solid	37.3	19.0	8.5	28.9	10.6	52.1
OpenSora	Fluid-Fluid	12.7	27.3	7.3	5.5	20.0	67.3
	Solid-Fluid	30.1	21.9	7.5	22.6	14.4	55.5
	Solid-Solid	7.7	23.8	1.4	6.3	22.4	69.9
VideoCrafter2	Fluid-Fluid	69.1	43.6	32.7	36.4	10.9	20.0
	Solid-Fluid	57.5	41.8	27.4	30.1	14.4	28.1
	Solid-Solid	31.5	23.8	4.9	26.6	18.9	49.7
SVD-T2I2V	Fluid-Fluid	58.2	34.5	18.2	40.0	16.4	25.5
	Solid-Fluid	52.7	32.9	17.1	35.6	15.8	25.5
	Solid-Solid	25.9	27.3	4.2	21.7	23.1	51.0
ZeroScope	Fluid-Fluid	36.4	47.3	20.0	16.4	27.3	36.4
	Solid-Fluid	40.4	37.0	14.4	26.0	22.6	37.0
	Solid-Solid	17.5	22.4	6.3	11.2	16.1	66.4
Closed Models							
Gen-2	Fluid-Fluid	37.7	26.4	15.1	22.6	11.3	50.9
	Solid-Fluid	38.5	18.5	8.1	30.4	10.4	51.1
	Solid-Solid	8.9	37.1	4.0	4.8	33.1	58.1
Lumiere-T2V	Fluid-Fluid	45.4	34.5	9.1	36.4	25.5	29.1
	Solid-Fluid	47.2	26.0	9.6	37.7	16.4	36.3
	Solid-Solid	26.5	27.3	8.4	18.2	18.9	54.5
Lumiere-T2I2V	Fluid-Fluid	49.5	21.8	10.9	38.2	10.9	40.0
	Solid-Fluid	59.6	26.0	17.1	42.5	8.9	31.5
	Solid-Solid	37.1	25.2	8.4	28.7	16.8	46.2
Pika	Fluid-Fluid	68.0	58.0	44.0	24.0	14.0	18.0
	Solid-Fluid	46.5	27.9	16.3	30.2	11.6	41.9
	Solid-Solid	24.8	36.8	13.6	11.2	23.2	52.0

Table 7: Fine-grained performance of T2V models for the complexity of the captions using human evaluation. Ideally, we want the T2V models to achieve a high score on the *SA = 1 and PC = 1* metric while reduce the score on the *SA=0 and PC=0*, *SA=1 and PC=0*, and *SA=0 and PC=1* metrics.

Source	Category	SA (%)	PC (%)	SA=1 and PC=1 (%)	SA=1 and PC=0 (%)	SA=0 and PC=1 (%)	SA=0 and PC=0 (%)
Open Models							
LaVIE	EASY	51.9	31.2	19.6	32.3	11.6	36.5
	HARD	44.8	24.0	11.0	33.8	13.0	42.2
OpenSora	EASY	20.1	25.4	4.8	15.3	20.6	59.3
	HARD	15.5	21.3	5.2	10.3	16.1	68.4
VideoCrafter2	EASY	53.4	38.1	21.2	32.3	16.9	29.6
	HARD	42.6	30.3	16.1	26.5	14.2	43.2
SVD-T2I2V	EASY	42.0	38.0	16.0	25.0	21.0	37.0
	HARD	43.0	23.0	6.0	37.0	16.0	41.0
ZeroScope	EASY	32.3	33.9	13.8	18.5	20.1	47.6
	HARD	27.7	31.0	9.7	18.1	21.3	51.0
Closed Models							
Gen-2	EASY	26.6	31.8	10.4	16.2	21.4	52.0
	HARD	26.6	21.6	4.3	22.3	17.3	56.1
Lumiere-T2V	EASY	38.6	34.9	11.1	27.5	23.8	37.6
	HARD	38.1	19.3	6.5	31.6	12.9	49.0
Lumiere-T2I2V	EASY	56.6	29.1	16.4	40.2	12.7	30.7
	HARD	38.7	20.0	7.7	31.0	12.3	49.0
Pika	EASY	45.7	39.9	23.7	22.0	16.2	38.2
	HARD	35.1	32.1	14.5	20.6	17.6	47.3

Semantic adherence:

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

Instruction (I): *[V] Does this video entail the description [T]?*
Response (R): ***Yes*** or ***No***

Figure 9: Template used assessing semantic adherence for a generated video.

Physical Commonsense:

Given: V (Video)

Instruction (I): *[V] Does this video follow physical laws?*
Response (R): ***Yes*** or ***No***

Figure 10: Template for assessing physical commonsense. We note that the physical commonsense is independent of the conditioning caption. Hence, it is not present in this template.

M Training Details for VIDEOCON-PHYSICS

To create VIDEOCON-PHYSICS, we use low-rank adaptation (LoRA) [31] of the VIDEOCON applied to all the layers of the attention blocks including QKVO, gate, up and down projection matrices. We set the LoRA $r = 32$ and $\alpha = 32$ and dropout = 0.05. The finetuning is performed for 5 epochs using Adam [35] optimizer with a linear warmup of 50 steps followed by linear decay. Similar to [3], we chose the peak learning rate as $1e-4$. We utilized 2 A6000 GPUs with the total batch size of 32. In addition, we finetune our model with 32 frames in the video and the frames are resized to 224×224 by image processor. Similar to [49, 90], we create 32 segments of the video, and sample the middle frame for each segment.

N VIDEOCON-PHYSICS Generalizes to Unseen Generative Models

To assess performance on an unseen video distribution, we train an ablated version of VIDEOCON-PHYSICS on a restricted set of video data. Specifically, we train VIDEOCON-PHYSICS on human annotations acquired from VideoCrafter2, ZeroScope, LaVIE, OpenSora, SVD-T2I2V, and Gen-2, and evaluate it on unseen videos from Lumiere-T2V, Lumiere-T2I2V, and Pika generated for the testing captions. We compare the performance of the zeroshot VIDEOCON and VIDEOCON-PHYSICS in Table 9. We find that VIDEOCON-PHYSICS outperforms VIDEOCON by 15 points and 15 points on semantic adherence and physical commonsense judgement, respectively. This highlights that VIDEOCON-PHYSICS can judge semantic adherence and physical commonsense as new T2V generative models are released.

O Applications of VIDEOCON-PHYSICS

In this work, we propose VIDEOCON-PHYSICS, an auto-evaluator that judges the semantic adherence and physical commonsense of the generated videos for a given caption. Here, we describe the potential usecases of the model for future work.

Video Generative Model Selection: The ability to perform model verification on downstream tasks cheaply and reliably is critical. In this regard, the model builders can utilize VIDEOCON-PHYSICS to evaluate their candidate models on the VIDEOPHY dataset at scale. The top candidate models can then be evaluated with the human workers for more accurate evaluation.

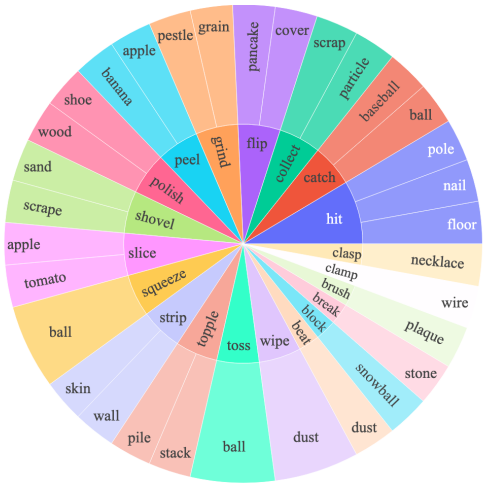


Figure 11: Top 20 most frequently occurring verbs (inner circle) and their top 4 direct nouns (outer circle) in our curated captions that consists of interaction between solid-solid states of matter.

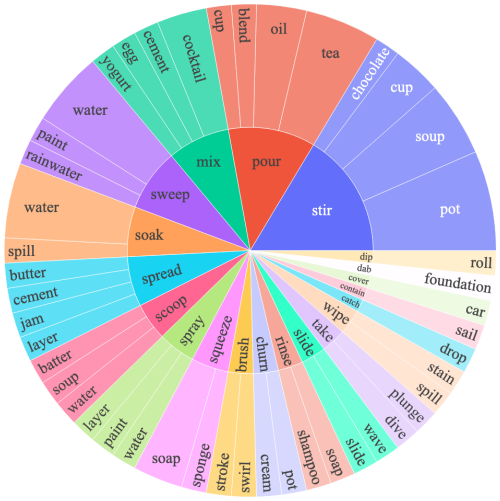


Figure 12: Top 20 most frequently occurring verbs (inner circle) and their top 4 direct nouns (outer circle) in our curated captions that consists of interaction between solid-fluid states of matter.

Data Filtering: With the advent of foundation models that are trained on the internet data, high-quality filtering has emerged as a crucial step in the pipeline [27, 95]. Here, the data builders can utilize VIDEOCON-PHYSICS to filter low-quality video-text data that lacks in semantic adherence and physical commonsense.

Post-training: Recently, aligning the generative models with human or AI feedback has become pivotal for high-quality generations [68, 62, 5, 79, 43]. Here, the post-training pipeline of the video generative models can leverage the VIDEOCON-PHYSICS model as an reward model that provides feedback to the model generated content. Subsequently, this feedback can be utilized to refine the model for better generations.

P More Qualitative Examples across Different Models

Here we compare results from VideoCrafter2 with results from other models. We find that LaVIE generates videos that show unnatural motions for the objects. Moreover, the videos tend to possess fast and vibrant dynamics. In addition, we observe that ZeroScope is prone to penetration issues as objects can be mixed with each other. Further, we find that Gen-2 does not understand gravity very

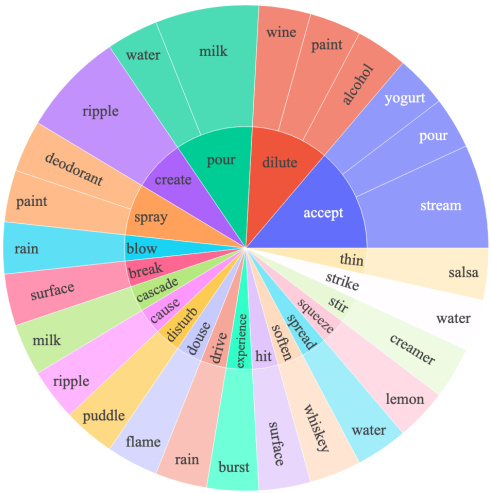


Figure 13: Top 20 most frequently occurring verbs (inner circle) and their top 4 direct nouns (outer circle) in curated captions that consists of interaction between fluid-fluid states of matter.

Table 8: Inference details for models in our testbed. Here, NA indicates that the information is not available for the closed models.

Model	Resolution	# of Video Frames	Guidance Scale	Sampling Steps	Noise Scheduler
Open Models					
ZeroScope	320 × 576	32	9	50	DPMSolverMultiStep [47]
VideoCrafter2	320 × 512	32	12	50	DDIM [73]
LaVIE	320 × 512	32	7.5	50	DDPM [29]
OpenSora	240 × 426	32	7	100	IDDPM [51]
SVD-T2I2V	1024 × 576	25	(1, 3)	25	EulerDiscrete [33]
Closed Models					
Lumiere-T2V	1024 × 1024	80	8	256	NA
Lumiere-T2I2V	1024 × 1024	80	6	256	NA
Gen-2	720 × 1280	32	8.5	100	NA
Pika	640 × 1088	72	12	NA	NA

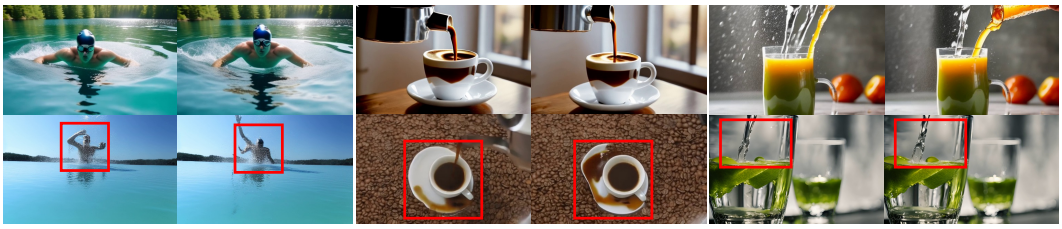
well, as objects that should be falling can be either static or even floating upwards, for instance, in Figure 14 (c).

Q More Qualitative Examples of Poor Physical Commonsense

We present more examples from each generative model where one or more physical laws are violated in Figure 15 - Figure 23.

Table 9: **Performance of VIDEOCON-PHYSICS on unseen generative model.** We train an ablated version of VIDEOCON-PHYSICS and find that it outperforms the baseline in the semantic adherence (SA) and physical commonsense (PC) judgment averaged over three unseen video models on the testing prompts.

Method	SA	PC
VideoCon [3]	64	57
VIDEOCON-PHYSICS (Ours)	79	72



(a) v. LaVIE. *A swimmer dives into a crystal-clear lake.*

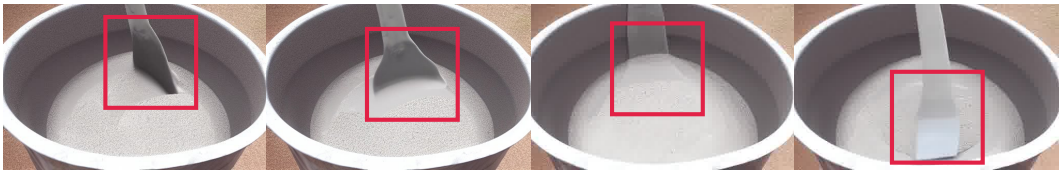
(b) v. ZeroScope. *A coffee pot pours a morning cup of joe.*

(c) v. Gen-2. *Water streams into fresh juice.*

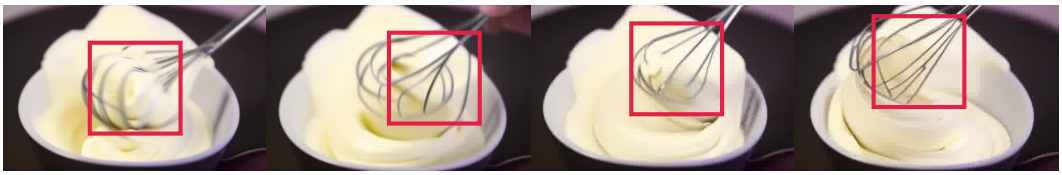
Figure 14: **Qualitative comparison of VideoCrafter2 with other models.** Here, the top row presents the generated videos from the VideoCrafter2 model. (a) LaVIE generated unnatural and inconsistent arm motions here. (b) The plate is deformed and penetrates the coffee beans beneath in the ZeroScope-generated video. (c) Gen-2 is prone to gravity issues. Here the stream of water is floating upwards, instead of flowing downwards.

R Examples from diverse states of matter and complexity

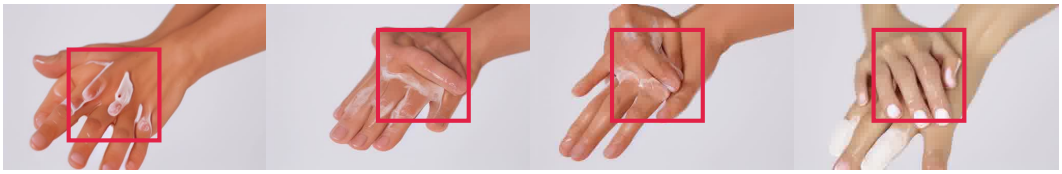
We present a few qualitative examples highlighting instances of good physical commonsense and bad physical commonsense in Figure 24-Figure 26.



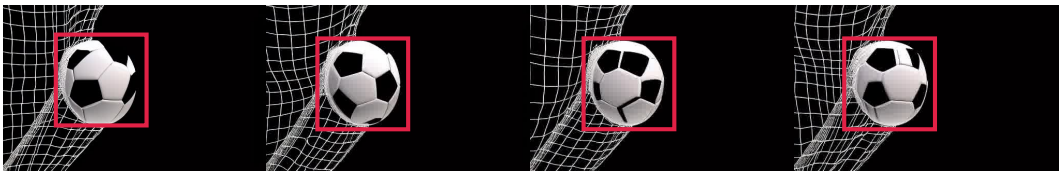
(a) *A paddle mixes wet cement in a bucket*



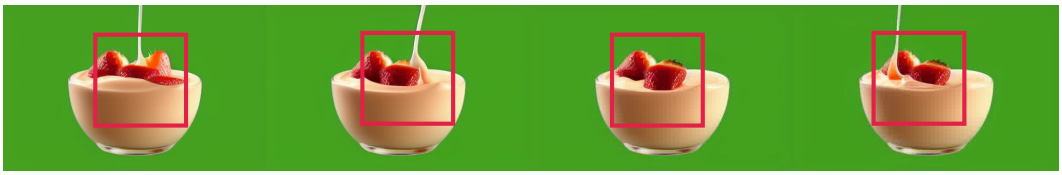
(b) *A whisk whips cream to a perfect fluffy consistency*



(c) *Hands rub luscious lotion on dry skin*



(d) *The net catches the fast-moving soccer ball*

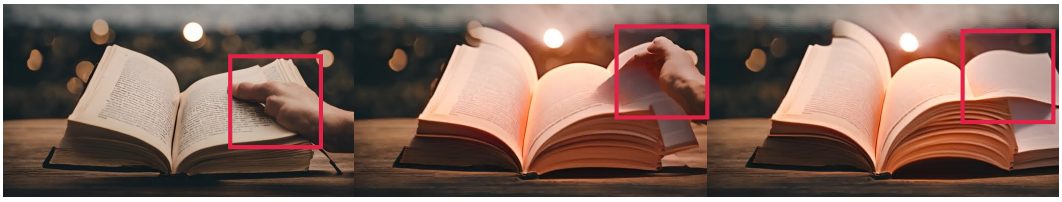


(e) *Yogurt merging with strawberry puree*

Figure 15: Unphysical Generated Examples of LaVIE. (a) Solid Constitutive Laws Violation: the metal spoon should not deform; Nonphysical Penetration: the spoon unnaturally passes through the liquid. (b) Solid Constitutive Laws Violation: the whisk exhibits abnormal shape deformation. (c) Solid Constitutive Laws Violation: the two hands show abnormal shape deformation; Nonphysical Penetration: fingers penetrate each other; Conservation of Mass Violation: the geometry (plus texture) of the two hands are inconsistent over time. (d) Conservation of Mass Violation: the geometry (plus texture) of the soccer is inconsistent over time; Newton's Second Law Violation: the soccer does not fall under gravity. (e) Conservation of Mass Violation: the volume of yogurt in the cup does not increase as more yogurt is added.

(a) *A blender spins, mixing squeezed juice within it*

(b) *A teaspoon stirs sugar into a cup of coffee*

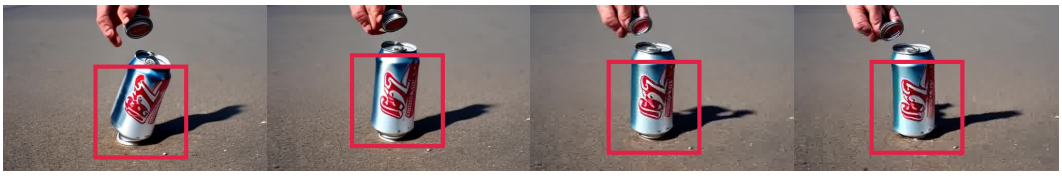


(c) *Hand flipping open book cover*

(d) *Soap washes grime off dirty hands*

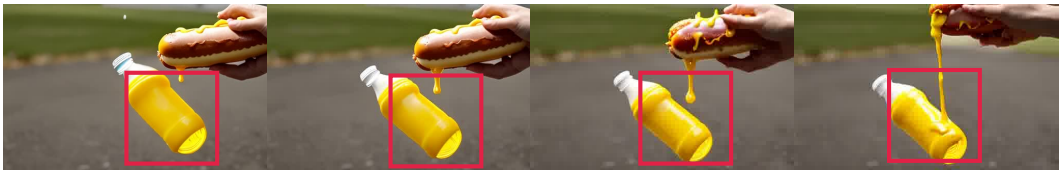
(e) *Water pouring from a watering can onto plants*

Figure 16: Unphysical Generated Examples of Gen-2. (a) Conservation of Mass Violation: the volume of juice in the blender increases over time without new substances being added. (b) Solid Constitutive Laws Violation: the metal spoon should not deform. (c) Conservation of Mass Violation: the volume of the book increases over time; Nonphysical Penetration: the fingers pass through the book. (d) Nonphysical Penetration: fingers penetrate into each other. (e) Newton's Second Law Violation: the flowing water appears to be static, ignoring the effect of gravity.

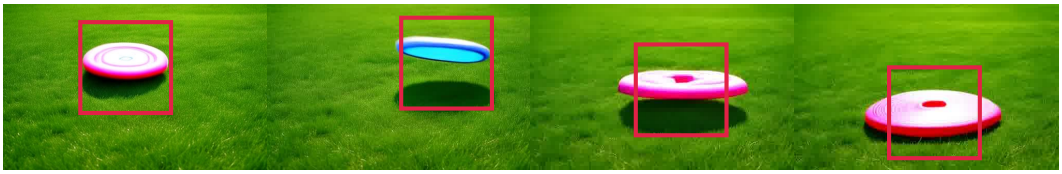


(a) *A foot crushing an empty soda can*

(b) *A spinning wheel sprays muddy water*



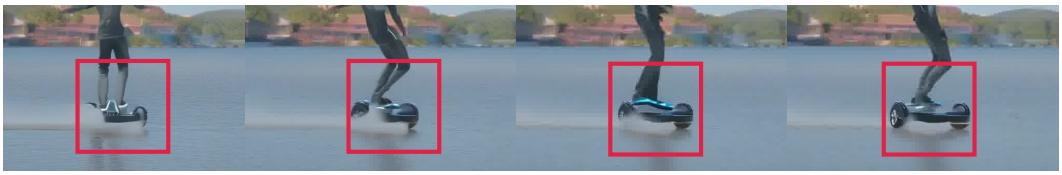
(c) *Mustard squirting out of a plastic bottle onto a hotdog*



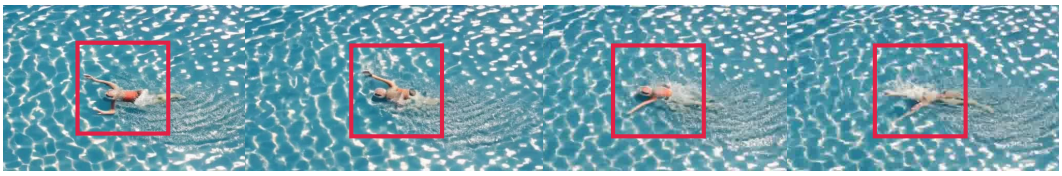
(d) *Plastic frisbee lands on a lush grass lawn*

(e) *Pouring milk into still tea*

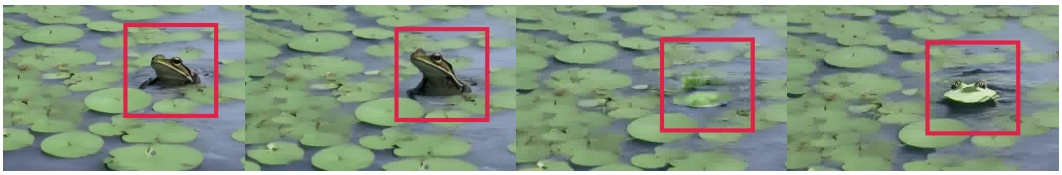
Figure 17: Unphysical Generated Examples of VideoCrafter2. (a) Newton's Second Law Violation: the metal can deforms without being pressed. (b) Newton's Second Law Violation: Water splashes while the rolling wheel remains static. (c) Newton's Second Law Violation: the bottle floats in the air, ignoring the effect of gravity; Fluid Constitutive Law Violation: the dripping and flowing of mustard are unnatural. (d) Conservation of Mass Violation: the geometry (plus texture) of the frisbee is not consistent over time. (e) Conservation of Mass Violation: the total volume of milk in the glass does not increase as more milk is poured into; Nonphysical Penetration: milk penetrates the glass.



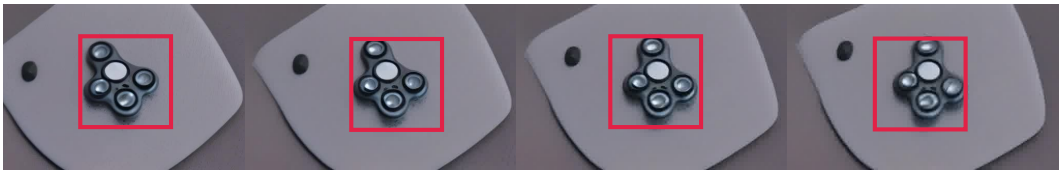
(a) *A futuristic hoverboard hovers just above the water*



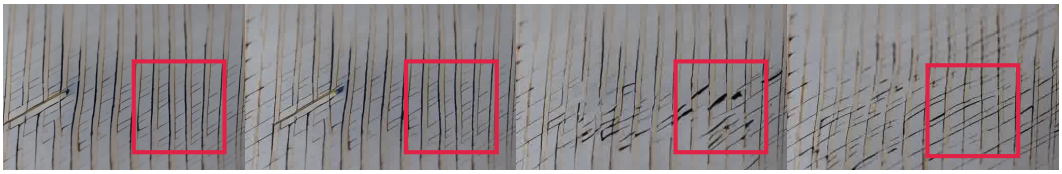
(b) *A swimmer splashing in the sea water*



(c) *Frog leaping from one lilypad to another*



(d) *Plastic fidget spinner rotating on rubber mat*

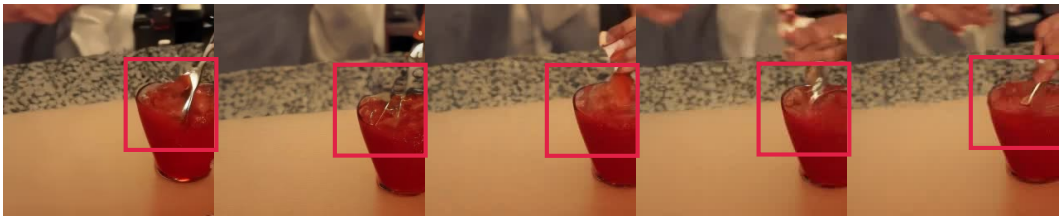


(e) *The eraser rubs against the paper, removing pencil marks*

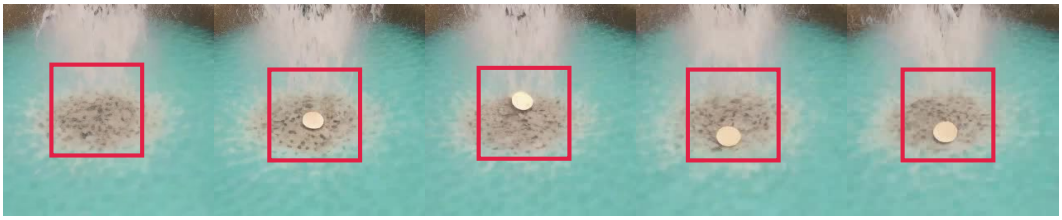
Figure 18: Unphysical Generated Examples of ZeroScope. (a) Newton's Second Law Violation: the motion of the hoverboard does not satisfy the momentum equation. (b) Newton's Second Law Violation: the motion of the arm of the swimmer is unnatural. (c) Conservation of Mass Violation: the geometry (texture) of the frog is inconsistent over time. (d) Newton's First Law Violation: the velocity of the fidget spinner changes despite being in a balanced state. (e) Solid Constitutive Laws Violation: the paper is torn apart without external forces but recovers later.

(a) *A blender spins, mixing squeezed juice within it*

(b) *A car gliding over a road slick with rainwater*



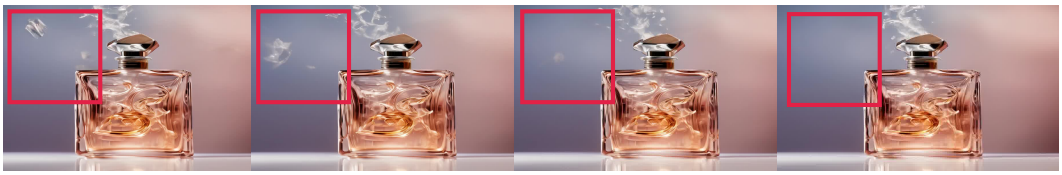
(c) *A shaker mixes a delightful cocktail at the bar*



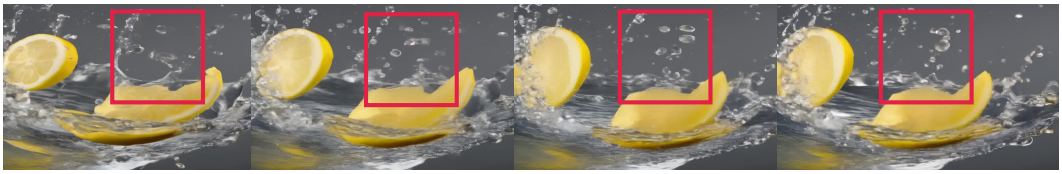
(d) *A shiny coin takes a dive into a clear water fountain*

(e) *A stroller wheels through a large puddle*

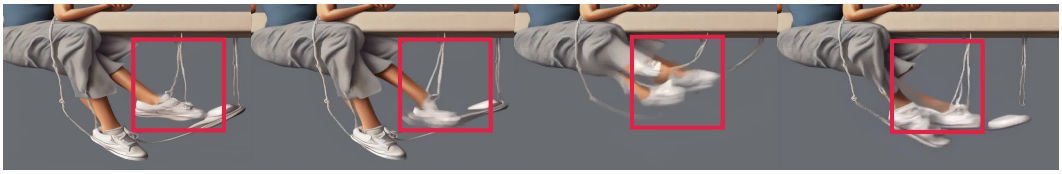
Figure 19: Unphysical Generated Examples of OpenSora. (a) Solid Constitutive Laws Violation: the metal blender should not deform. (b) Newton's Second Law Violation: the car moves backward, violating the momentum equation (c) Solid Constitutive Laws Violation: the metal spoon deforms when stirring the cocktail. (d) Newton's First Law Violation: the coin moves on the ground back and forth without horizontal forces. (e) Conservation of Mass Violation: the left rear wheel disappears over time.



(a) *A perfume bottle spritzes fragrance into the air*

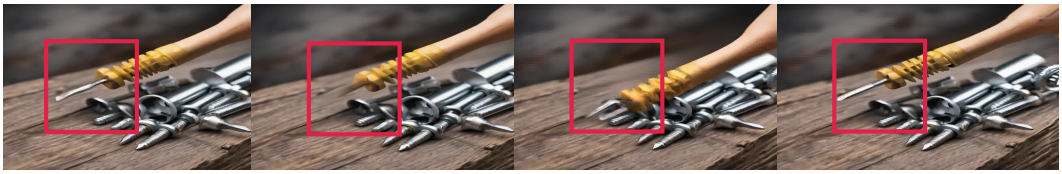


(b) *Lemon juice drops splash into water*



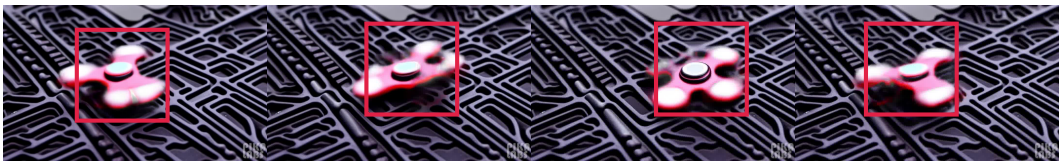
(c) *Loose sneaker swings on dangling foot*

(d) *Detergent flowing into a bucket of water*

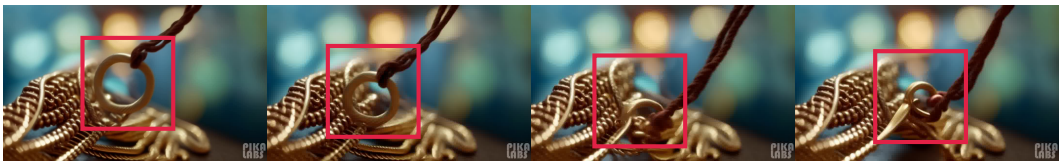


(e) *The screwdriver tightens the metal screw in the wood*

Figure 20: Unphysical Generated Examples of SVD-T2I2V. (a) Newton's Second Law Violation: the perfume spreads back and forth, violating the momentum equation. (b) Newton's Second Law Violation: the water drops float in the air, ignoring gravity. (c) Solid Constitutive Laws Violation: the leg exhibits unnatural deformation. (d) Newton's Second Law Violation: the water flows upward into the air without external forces. (e) Solid Constitutive Laws Violation: the screwdriver head deforms unnaturally.

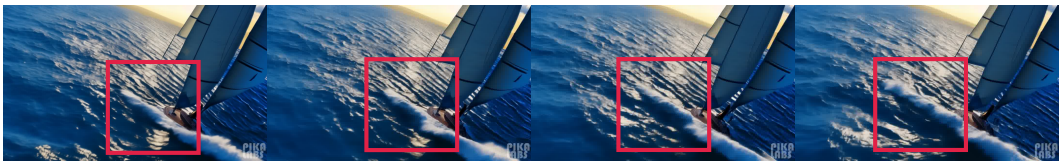


(a) *Plastic fidget spinner rotating on rubber mat*



(b) *Clasping a necklace around a neck*

(c) *A whisk churns heavy cream into whipped cream*

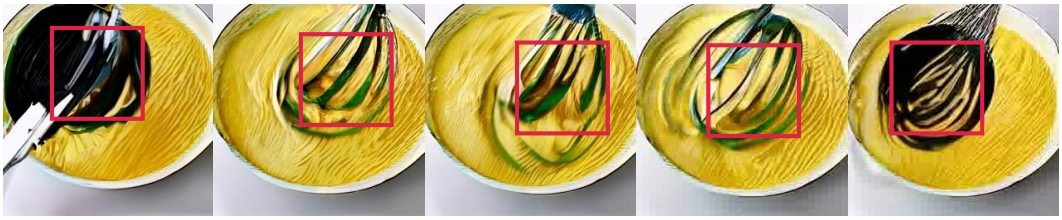


(d) *A sailboat cuts through the choppy sea waves*

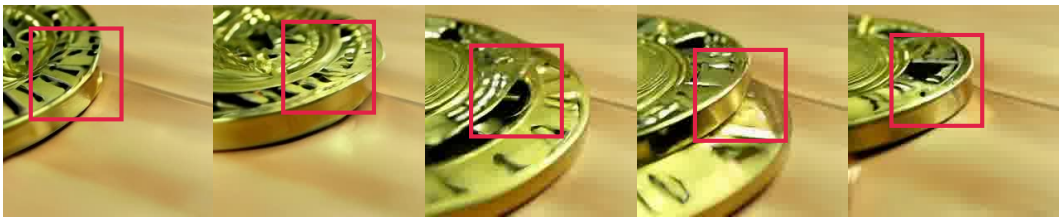
(e) *A diver plunges headlong into a sparkling pool*

Figure 21: Unphysical Generated Examples of Pika. (a) Solid Constitutive Laws Violation: the fidget spinner should not deform. (b) Solid Constitutive Laws Violation: the necklace should not deform. (c) Conservation of Mass Violation: the volume of cream increases over time without additional input. (d) Fluid constitutive Law Violation: unnatural waves on the sea surface. (e) Solid Constitutive Law Violation: one diving shoe splits into two and detaches from the feet.

(a) *A spoon stirs a pot of vegetable soup*



(b) *A whisk spins in the egg mixture, mixing it thoroughly*



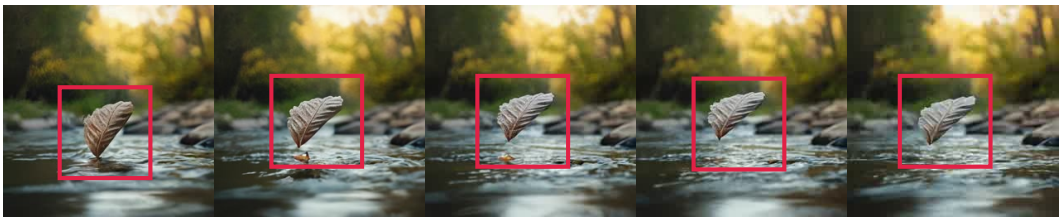
(c) *Coin spins rapidly on a wooden table*

(d) *Squeezing lemon drops into warm tea*

(e) *Tea accepts stream of milk*

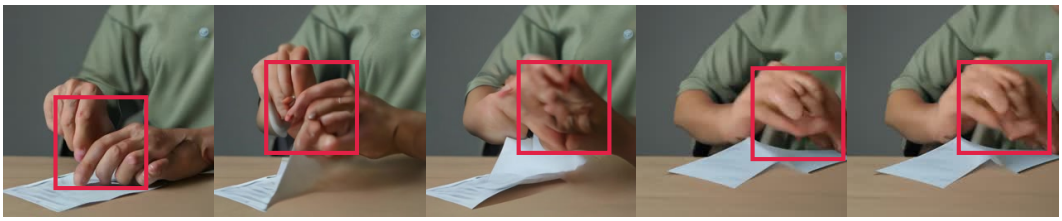
Figure 22: Unphysical Generated Examples of Lumiere-T2V. (a) Conservation of Mass Violation: the vegetable appears on the spoon out of nowhere. (b) Solid Constitutive Laws Violation: the whisk should not deform. (c) Solid Constitutive Laws Violation: the coin splits into two and then merges back into one. (d) Solid Constitutive Laws Violation: the lemon shows an unnatural appearance change; Fluid Constitutive Laws Violation: the lemon juice appears like static glue. (e) Nonphysical Penetration: the tea flows through the cup.

(a) *A drum vibrating from the beating stick*



(b) *A leaf falls delicately into a slow-moving river*

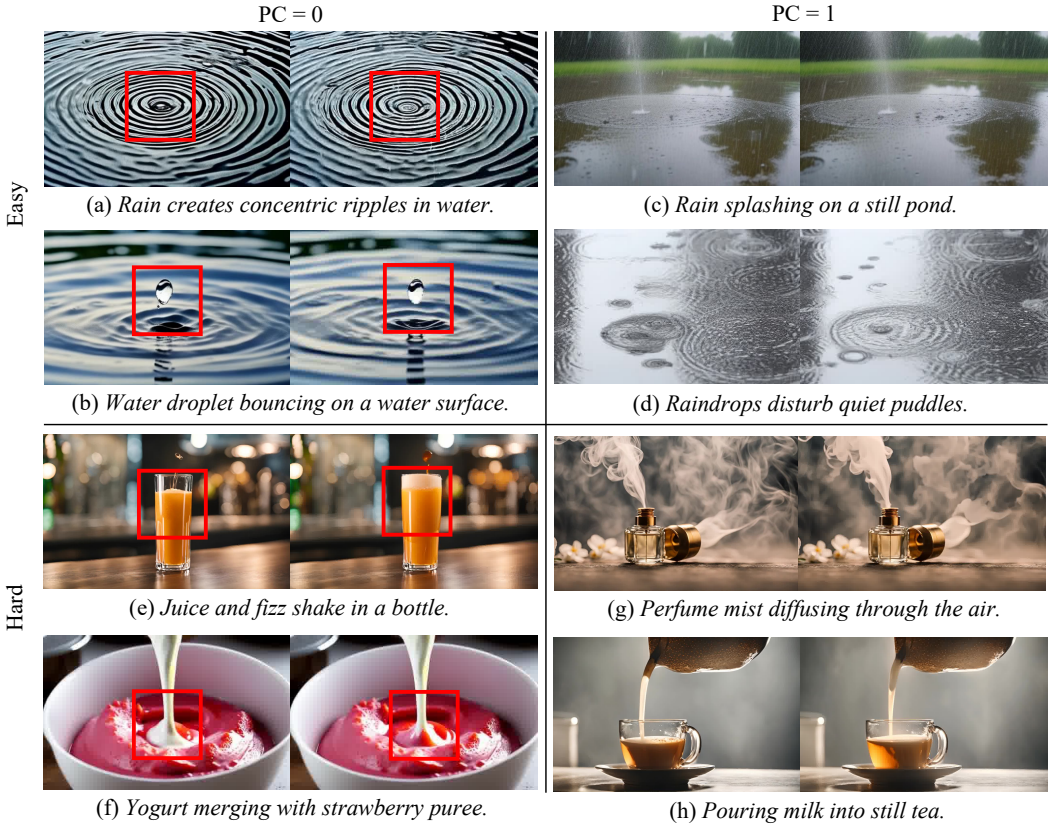
(c) *A wooden spoon stirring soup in a pot*



(d) *Hand folds the paper*

(e) *Skateboard glides on the pavement*

Figure 23: Unphysical Generated Examples of Lumiere-T2I2V. (a) Solid Constitutive Laws Violation: the drum stick head should not deform (b) Newton's Second Law Violation: the leaf floats in the air, ignoring gravity. (c) Conservation of Mass Violation: the vegetable appears on the spoon out of nowhere. (d) Nonphysical Penetration: hands penetrate each other. (e) Solid Constitutive Laws Violation: one leg on the skateboard transforms into a person.



(a) *Rain creates concentric ripples in water.*

(c) *Rain splashing on a still pond.*

(b) *Water droplet bouncing on a water surface.*

(d) *Raindrops disturb quiet puddles.*

(e) *Juice and fizz shake in a bottle.*

(g) *Perfume mist diffusing through the air.*

(f) *Yogurt merging with strawberry puree.*

(h) *Pouring milk into still tea.*

Figure 24: Qualitative examples in the fluid-fluid category. Videos in the left column have a majority PC score of 0, while videos in the right column have a majority PC score of 1. (a) The central ripple does not vanish even in absence of raindrops. (b) The water droplet is floating upwards, defying gravity. (e) The total volume of juice is increasing. (f) The color of the yogurt is not consistent over time.

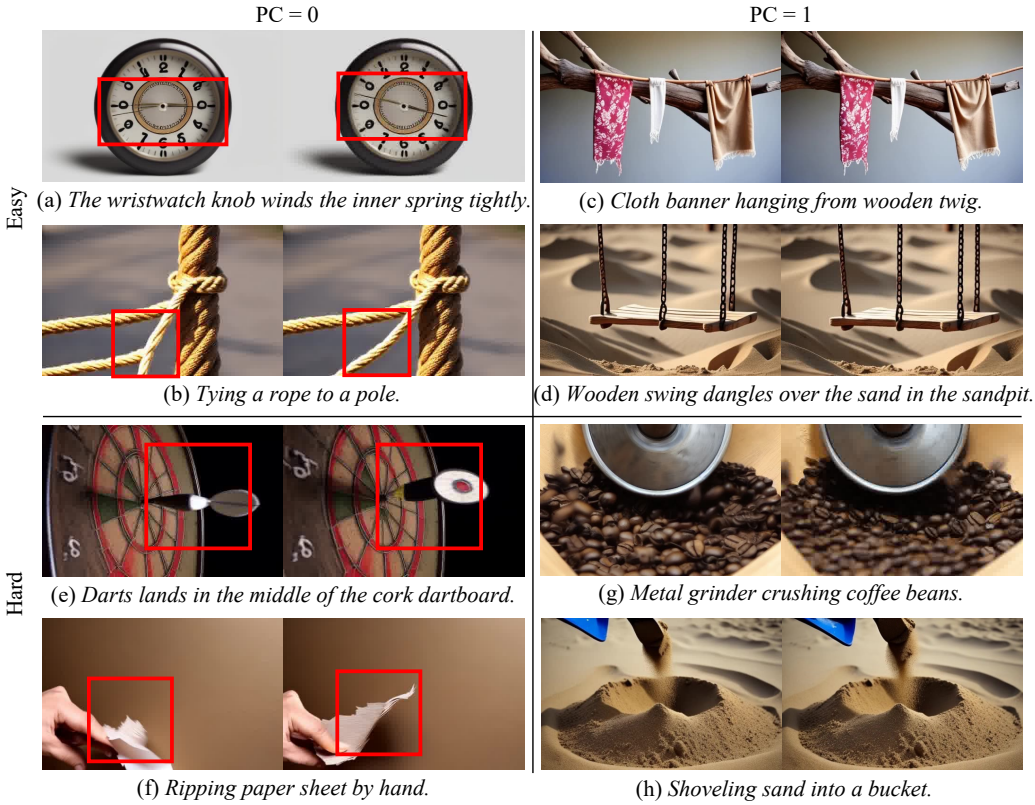
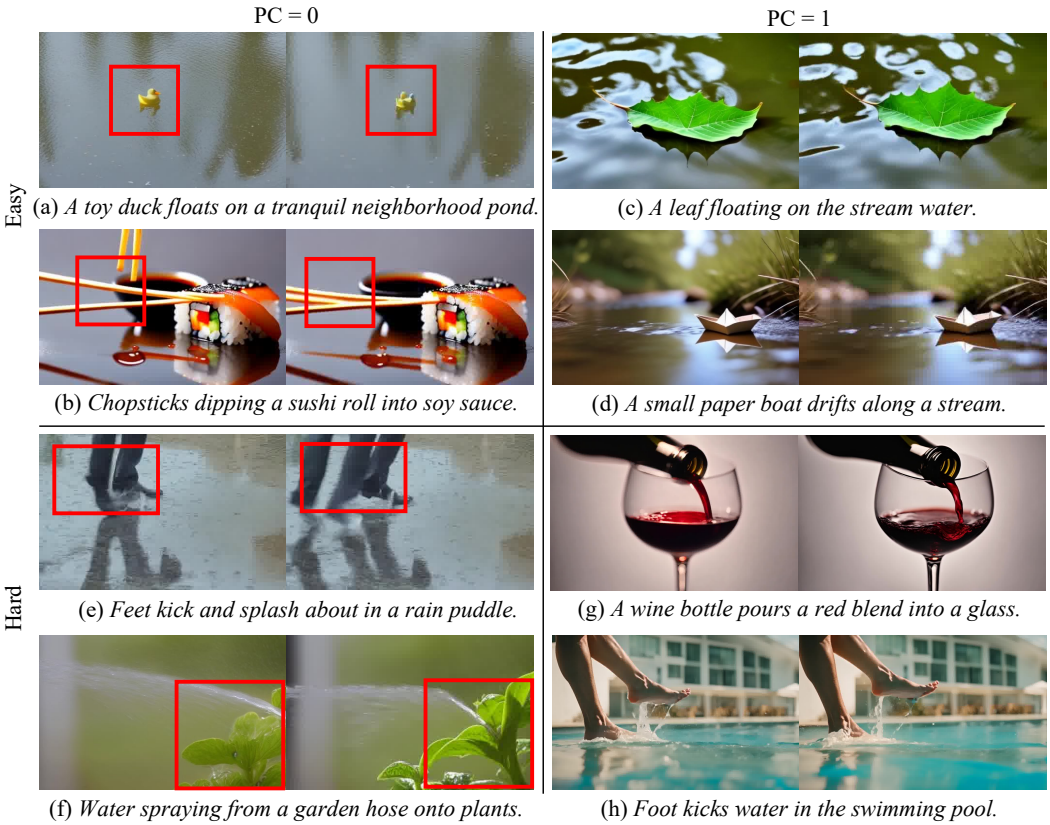


Figure 25: Qualitative examples in the solid-solid category. Videos in the left column have a majority PC score of 0, while videos in the right column have a majority PC score of 1. (a) The hands of the clock have illogical motion. (b) One piece of the robe disappears. (e) The geometry and texture of the dart are not consistent over time. (f) The total volume of the sheet of paper is not consistent over time.



(a) *A toy duck floats on a tranquil neighborhood pond.*
(c) *A leaf floating on the stream water.*
(b) *Chopsticks dipping a sushi roll into soy sauce.*
(d) *A small paper boat drifts along a stream.*
(e) *Feet kick and splash about in a rain puddle.*
(g) *A wine bottle pours a red blend into a glass.*
(f) *Water spraying from a garden hose onto plants.*
(h) *Foot kicks water in the swimming pool.*

Figure 26: Qualitative examples in the fluid-fluid category. Videos in the left column have a majority PC score of 0, while videos in the right column have a majority PC score of 1. (a) The geometry and color of the duck head changes over time. (b) One chopstick appears from nowhere. (e) One leg appears from nowhere. (f) The geometry and texture of the leaves are not consistent over time.