TOWARDS WORLD SIMULATOR: CRAFTING PHYSICAL COMMONSENSE-BASED BENCHMARK FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-video (T2V) models like Sora have made significant strides in visualizing complex prompts, which is increasingly viewed as a promising path towards constructing the universal world simulator. Cognitive psychologists believe that the foundation for achieving this goal is the ability to understand intuitive physics. However, the capacity of these models to accurately represent intuitive physics remains largely unexplored. To bridge this gap, we introduce *PhyGenBench*, a comprehensive **Phy**sics **Gen**eration **Bench**mark designed to evaluate physical commonsense correctness in T2V generation. *PhyGenBench* comprises 160 carefully crafted prompts across 27 distinct physical laws, spanning four fundamental domains, which could comprehensively assess models' understanding of physical commonsense. Alongside *PhyGenBench*, we propose a novel evaluation framework called *PhyGenEval*. This framework employs a hierarchical evaluation structure utilizing appropriate advanced vision-language models and large language models to assess physical commonsense. Through *PhyGenBench* and *PhyGenEval*, we can conduct large-scale automated assessments of T2V models' understanding of physical commonsense, which aligns closely with human feedback. Our evaluation results and in-depth analysis demonstrate that current models struggle to generate videos that comply with physical commonsense. Moreover, simply scaling up models or employing prompt engineering techniques is insufficient to fully address the challenges presented by *PhyGenBench* (e.g., dynamic physical phenomenons). We hope this study will inspire the community to prioritize the learning of physical commonsense in these models beyond entertainment applications. We release the data and codes at `https://github.com/PhyGenBench/PhyGenBench`

1 INTRODUCTION

Text-to-video (T2V) models such as Sora have made significant strides in visualizing complex ideas and scenes based on textual input (Yang et al., 2024; Wang et al., 2023). These advancements are increasingly viewed as a promising path towards constructing universal simulators of the physical world, which holds immense promise for video generation (Zhu et al., 2024), autonomous driving (Gao et al., 2024), and the development of embodied agents (Mazzaglia et al., 2024). Cognitive psychology posits that intuitive physics, which is demonstrated even by human infants (Wood et al., 2024; Battaglia et al., 2013), is essential for achieving this goal. Intuitive physics emphasizes rendered scenes should be visually and interactively natural to humans, rather than adhere to strict physical accuracy. Consequently, on the path towards developing a world simulator (Xiang et al., 2024), video generation should first be capable of accurately reproducing simple yet fundamental physical phenomenons. However, even state-of-the-art models trained on vast resources (Tan et al., 2024) encounter difficulties in correctly generating seemingly trivial physical phenomenons, as depicted in Figure 1, the model fails to understand that the stone should sink in water. This clear pitfall shows a substantial gap between current video generation models' and human's understanding of basic physics. It reveals how far these models are from being true world simulators.

Given this context, it becomes important to assess the extent to which current T2V models can capture intuitive physics in their generated outputs. This requires the development of comprehensive

Figure 1: Samples of videos generated by Kling or Gen-3 in *PhyGenBench* with 4 different aspects. The results show that current T2V models struggle to generate videos that align with physical commonsense (e.g., the lack of a plane's reflection in water in the first video of the second row).

evaluation frameworks that beyond traditional metrics. While numerous Text-to-Video (T2V) evaluation benchmarks have emerged (Sun et al., 2024; Huang et al., 2024), they primarily focus on various qualities of generated videos (e.g., motion smoothness, background consistency) or spatial relationships, failing to address the critical issue of whether the generated videos adhere to fundamental physical laws. Although some studies have explored the alignment of generated videos with dynamic motions naturalness (Bansal et al., 2024), their benchmarks fail to succinctly capture fundamental physical laws or propose sufficiently robust evaluation methods. Therefore, the development of benchmarks and evaluation methodologies specifically tailored to assess intuitive physics in generated videos remains a critical yet largely unexplored frontier.

There are two challenges impeding the evaluation of physical commonsense in T2V models. On one hand, there is a lack of benchmarks focused on evaluating physical commonsense. This requires selecting semantically simple physical phenomenons that exhibit clear physical phenomena, allowing for accurate assessment by either humans or machines. On the other hand, there is a lack of corresponding evaluation metrics. Traditional metrics like FVD (Unterthiner et al., 2018) exhibit limitations in detecting implausible motions (Brooks et al., 2022) and necessitate reference videos, which are often challenging to procure for novel scenes. Recent studies have used video-based VLMs for comprehensive video evaluation (He et al., 2024b; Sun et al., 2024). However, they often struggle to correctly assess physical commonsense. This limitation stems from their inadequate understanding of physical laws (Jassim et al., 2023) and the fact that these methods are not specifically designed to evaluate physical laws.

To address these challenges, we propose *PhyGenBench* and *PhyGenEval* to automate the evaluation of physical commonsense understanding capability from T2V models. *PhyGenBench* is designed to evaluate physical commonsense based on fundamental physical laws in text-to-video generation. Inspired by (Halliday et al., 2013), we categorize physical commonsense in the world into four main areas: mechanics, optics, thermal, and material properties. Then principle physical laws and easily observable physical phenomenons are identified for each category, resulting in comprehensive 27 physical laws and 160 validated prompts in the proposed benchmark. Through brainstorming, we construct prompts that easily reflect physical laws using sources like textbooks (Harjono et al., 2020). This process results in a comprehensive but simple set of prompts reflecting physical commonsense, which are sufficiently clear for evaluation. As shown in Figure 1, the correctness of physical commonsense in *PhyGenBench* can be observed through clear phenomena (e.g., *plane should have reflections in water*) On the other hand, benefiting from the simple yet clear physical

phenomena in *PhyGenBench* prompts, we can propose *PhyGenEval*, which is a novel video evaluation framework for assessing physical commonsense correctness in *PhyGenBench*. *PhyGenEval* first uses GPT-4o[1] to analyze physical laws in text, addressing the poor understanding of physical common sense in video-based VLMs. Moreover, considering that previous evaluation metrics did not specifically target physical correctness, we propose a three-tier hierarchical evaluation strategy for this aspect, transitioning from image-based to comprehensive video analysis: single image, multiple images, and full video stages. Each stage employs distinct VLMs along with custom instructions generated by GPT-4o to form judgments. By combining *PhyGenBench* and *PhyGenEval*, we can efficiently evaluate different T2V models' understanding of physical commonsense at scale, producing results highly consistent with human feedback.

The contributions of our work are three-fold. **i):** We proposed *PhyGenBench*, which compasses a wide range of clear physical phenomenons and explicit physical laws. This benchmark can comprehensively measure whether T2V models understand intuitive physics and indirectly assess their gap from world simulator capabilities **ii):** Along with the benchmark, we propose an automated evaluation framework - *PhyGenEval*, which overcomes the challenges of assessing the correctness of physical commonsense with other metrics and demonstrates high consistency with human feedback on *PhyGenBench*, enabling users to conduct large-scale automated testing of various T2V models. **iii):** We conduct extensive evaluations of popular T2V models, even the best-performing model, Gen-3, scores only 0.51. This indicates that current models are still far from functioning as world simulators. Based on our evaluation results, we conduct an in-depth analysis and discover that addressing issues such as dynamics is still challenging through prompt engineering or simply scaling up model. We hope this work inspires the community to focus on the learning of physical commonsense in T2V models, rather than merely using them as tools for entertainment.

2 Related work

2.1 Benchmarks for text-to-video generation

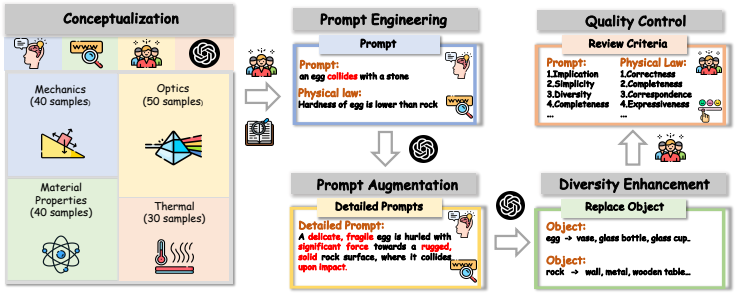
The rapid advancement of text-to-video (T2V) generation models has necessitated various benchmarks for accurate assessment. Traditional works in video generation, such as FVD (Unterthiner et al., 2018), rely on datasets like UCF-101 (Soomro, 2012) and Kinetics-400 (Kay et al., 2017), which are limited in scope. Recent benchmarks, including VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024c), aim to comprehensively evaluate general video quality across multiple dimensions. In contrast, some studies focus on fine-grained evaluation of text-to-video (T2V) models from specific aspects. For instance, T2V-CompBench (Sun et al., 2024) assesses compositional generation capabilities, while DEVIL (Liao et al., 2024) evaluates dynamic characteristics of generated videos. Although some research like VideoPhy (Bansal et al., 2024) efforts address the dynamic motions naturalness of video generation, their benchmarks fail to succinctly capture fundamental physical laws. Consequently, most existing works overlook this crucial aspect, which forms the foundation for realizing a world simulator. To address this gap, we introduce *PhyGenBench*, a benchmark designed to comprehensively measure T2V models' understanding of physical commonsense.

2.2 Evaluation metrics for text-to-video generation

Conventional approaches to video quality assessment often employ metrics such as FVD (Unterthiner et al., 2018) and IS (Salimans et al., 2016). However, the detection of unrealistic motions is difficult for them (Brooks et al., 2022), and FVD requires a reference video that is hard to obtain for novel scenes, making it challenging to evaluate the correctness of physical commonsense. Recent studies have explored the use of advanced vision-language models (VLMs) as evaluators. For instance, VideoScore (He et al., 2024b) leverages human feedback to train models for video quality assessment, while T2V-CompBench (Sun et al., 2024) utilizes powerful models like LLaVA (Liu et al., 2024a) to evaluate the correctness of spatial relationships. Although a few works demonstrate improved alignment with human judgments, they fall short in generalizing to assessments of physical commonsense correctness. To address this limitation, we introduce *PhyGenEval*, a novel

[1]The version is gpt4o-0806

(a) The overview of the PhyGenBench



(b) The construction pipeline of the PhyGenBench

Figure 2: (a) is the overview of the proposed *PhyGenBench*. (b) is the *PhyGenBench* data pipeline, which covers four physics categories. We select key physical laws and manually craft initial prompts that reflect the corresponding physical phenomena. GPT-4o adds details and enhances diversity by varying objects. After manual review, we obtain 160 T2V prompts.

method designed to evaluate physical commonsense correctness on *PhyGenBench*. We validate the efficacy of our approach through comprehensive human correlation studies.

3 PHYGENBENCH

Inspired by (Swartz, 1985), we first define the following terms: *"Physical Commonsense:"* Basic intuitive understanding of how physical objects and actions behave in everyday life; *"Physical Laws:"* Universal scientific principles that describe consistent behaviors in nature; *"Physical Phenomenon:"* Observable events or processes caused by the interaction of physical laws. The purpose of *PhyGenBench* is to evaluate whether T2V models understand physical commonsense, while each prompt in *PhyGenBench* presents a clear physical phenomenon and an underlying physical law.

Overview. As illustrated in Figure 2 (a), *PhyGenBench* encompasses four major categories of physical commonsense: *"Mechanics"*, *"Optics"*, *"Thermal"*, and *"Material Properties"*. It incorporates 27 physical phenomena with intrinsic physical laws reflected by the corresponding designed 160 prompts:

1. *"Mechanics"* covers 7 common mechanical laws: gravity, buoyancy, solid pressure, atmospheric pressure, elasticity, friction, and surface tension, with 40 validated prompts. For example, we use *"A piece of iron is gently placed on the surface of the water in a tank filled with water"* to test T2V model's understanding of Buoyancy, where the iron should sink due to its higher density compared to water.

2. *"Optics"* categorizes 6 aspects based on light phenomena: reflection, refraction, scattering, dispersion, interference & diffraction, and straight-line propagation, yielding 50 prompts. A prompt like *"a kite soaring above a smooth and tranquil pond"* is used to test reflection generation capability.

3. *"Thermal"* considers 6 phase transitions: Solidification, Melting, Liquefaction, Boiling, deposition, Sublimation, comprising 30 prompts. Inspired by ChronoMagicBench (Yuan et al., 2024), the vaporization (boiling) process is evaluated by the prompt *"a timelapse capturing the transformation of water as the temperature rapidly rises above* $100°C$*"*.

4. *"Material Properties"* includes 5 physical properties (color, hardness, solubility, combustibility, and flame reaction) and 3 chemical properties (acidity, redox potential, and dehydrating properties), resulting in 40 prompts. We reflect material properties, e.g., *"hardness"*, through the prompts with expected phenomena, e.g., *"an egg being hurled with significant force towards a rock"*, where the egg should break while the rock remains intact.

Multiple physical laws could be included in a single prompt, which may bring confusion to the evaluation of physical common sense in video generation, even for human annotators. To avoid this, we carefully curate prompts to ensure a one-to-one correspondence for each physical phenomenon it reflects, with clear physical law inside. By incorporating physical laws from four distinct physical categories, *PhyGenBench* offers a thorough assessment of current T2V models' understanding of physical commonsense.

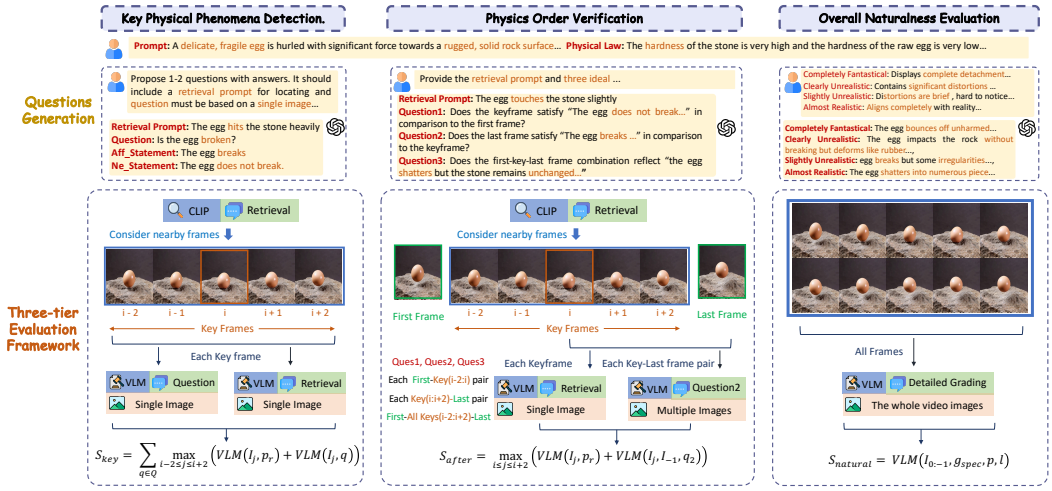
Benchmark Construction. As shown in Figure 2 (b), we develop a comprehensive approach to create *PhyGenBench*. The methodology encompasses five steps: **1) Conceptualization:** Following (Halliday et al., 2013), We first identify key physical commonsense from four major categories of physics. For each category, we select specific physical laws from textbooks (Harjono et al., 2020), which can be widely recognized and can be easily demonstrated through clear, observable physical phenomenon. **2) Prompt Engineering:** For each physical law, we manually craft the initial T2V prompts to clearly depict the underlying physical phenomenon **3) Prompt Augmentation:** To enhance the model's video generation capabilities, we augment the initial T2V prompts by adding additional details, such as more precise descriptions of objects and actions (Yang et al., 2024). This augmentation process is carefully designed to avoid revealing the expected physical phenomenon. **4) Diversity Enhancement:** Following T2V-CompBench (Sun et al., 2024), we employ GPT-4o to perform object substitution on the augmented prompts. This step increases the diversity of the benchmark. **5) Quality Control:** We conduct a thorough review of the prompts and their associated physical laws to ensure accuracy and relevance. Specifically, we ensure that the T2V prompts and corresponding physical laws are clear and accurate. We then randomly use the current T2V model to check if the prompts are simple enough for the model to generate semantically accurate videos. This methodology yields a robust and comprehensive benchmark for assessing T2V models' comprehension of physical commonsense, providing a valuable tool for advancing research in this domain. For more detailed information about the dataset, please refer to the Appendix B

4 PHYGENEVAL

PhyGenEval aims to assess whether the physical phenomena in the generated videos conform to the corresponding physical laws. To obtain a clear judgment, we decompose the evaluation into semantic alignment (SA) and physical commonsense alignment (PCA). While SA evaluates whether the semantic meaning inferred by the generated video and the input prompt are matched, PCA measures whether the evaluated physical laws are grounded in the videos. For example, for the scene *"an egg collides with a stone"*, SA requires a video containing the egg, the stone, and the collision action. PCA necessitates a video for the whole physical motions in which the egg hits a stone and then breaks, while the stone remains intact. Following (He et al., 2024b), we convert both SA and PCA to a four-point scale, as well as the human ratings.

4.1 SEMANTIC ALIGNMENT EVALUATION

Directly asking the Vision-Language Model(VLM) to align the semantic meaning between videos and input prompts are difficult, as prompts usually are mixed with semantic entities and physical phenomena, and the intermediate outcomes are subtly implied by the videos. For example, in a prompt like *"A timelapse captures the transformation of soup as the temperature rises above 100°C"*, a possible video generation would appear like *"The video shows a soup, but there is no transformation of the soup"*. To address the challenge, we first employ GPT-4o to extract object and action from the original text prompt, we then utilize GPT-4o to sequentially determine the presence of extracted objects in the video and verify the occurrence of specified actions. This decomposition provides more fine-grained captures and prevents the model from confusing semantic and physical correctness during evaluation. Experimental results demonstrate that our automated evaluation method aligns more



(a) Key Physical Phenomena Detection (b) Physics Order Verification (using Question 2 as an example) (c) Overall Naturalness Evaluation

Figure 3: An overview of the proposed *PhyGenEval*. *PhyGenEval* is divided into three parts: Key Physical Phenomena Detection, Physics Order Verification, and Overall Naturalness Evaluation. Each part uses an appropriate VLM in combination with physical-based customized questions generated by GPT-4o. The final score is the combined result of the three parts. For the example in the figure, the three-stage scores are 0, 1 (only $Question_1$ is correct), and 0. The final score is calculated as 0 according to Overall Score calculation in Section 4.2.

closely with human judgment and outperforms previous methods (He et al., 2024b; Sun et al., 2024) in *PhyGenBench* (Appendix C.1).

4.2 Physical Commonsense Evaluation

To evaluate physical correctness in the video, we evaluated multiple common evaluation metrics comparing human assessments[2]. Experimental results in Table 1 demonstrate that these methods struggle to generalize to the assessment of physical commonsense correctness on *PhyGenBench*, e.g., VideoScore (He et al., 2024b) has only a spearman correlation of 0.19 on *PhyGenBench*, which is most correlated with human assessments except *PhyGenEval*. We attribute it to the main factor: Directly using video-based VLMs fails to comprehend the embedded physical commonsense (Jassim et al., 2023), as current methods are not designed with physical commonsense as a foundation. To fully understand the physical commonsense in the video, there are three key factors need to solve: **i):** Physical processes typically exhibit clear key phenomena depicted by the input prompt (e.g., *"the egg breaks upon hitting the rock."*). It is necessary to identify these key physical phenomena and detect their presence in videos. **ii):** Physical processes are characterized by causality, manifested in the correct sequence of critical events(e.g., *"The egg touchs the rock first, then breaks."*). The correct sequence order validates the correctness of physical processes. **iii):** Physical processes need to possess overall naturalness, which represents the realistic of the overall process. To address these factors, we design a progressive strategy that starts with key physical phenomena, then moves through the sequence of several key phenomena, and finally evaluates the overall naturalness of the entire video. This hierarchical and refined approach reduces the difficulty compared to existing methods that directly uses VLMs to evaluate physical commonsense, enabling *PhyGenEval* to achieve results closely aligned with human judgements.

Key Physical Phenomena Detection. This stage aims to detect *whether the key physical phenomena occur in the video.* Here we define the key phenomena as an observable and distinctive occurrence (e.g., a specific frame) within a physical process that can directly reveal the corresponding physical law, like deformations or color changes. For each input prompt in *PhyGenBench*, we craft a retrieval prompt p_r and a set of physics-related questions Q, where the retrieval prompt is

[2]Annotators are asked to score the correctness of physical commonsense in the video. Details refer to Section 5 and Appendix D.1

used to locate the key phenomena frame, and physical-related questions are utilized to check whether the expected physics phenomena are present in the keyframe.

As illustrated in Figure 3 (a), we first obtained both Q and P_r by prompting GPT-4o with the input T2V prompt and corresponding physical law. Following (Hessel et al., 2021), a keyframe I_i from the video based on the retrieval prompt, where I_i is the i-th frame in the video. By using the keyframe, we define a confidence score of the key phenomena in the video:.

$$\mathrm{S}_{\mathrm{key}} = \sum_{q \in Q} \max_{i-2 \leq j \leq i+2} \left(\mathrm{VLM}(I_j, q) + \mathrm{VLM}(I_j, p_r) \right),$$

where $\mathrm{VLM}(I_j, q)$ reflects the presence of physical phenomena in I_j for each related question q from Q. $\mathrm{VLM}(I_j, p_r)$ checks whether I_j matches the retrieval prompt, which ensures key phenomena occur at the correct frame. Since videos may contain semantic errors, it's also important for determining if key physical phenomena occur (e.g., an egg shouldn't break in mid-air before hitting a rock). We consider adjacent 5 frames near the keyframe to enhance the robustness. For example, the egg may not be cracked just when it first contacts the stone. We instantiate VLM-based evaluator $\mathrm{VLM}(\cdot)$ with VQAScore (Lin et al., 2024), which has been shown promising evaluation results on visual question-answering.

Physics Order Verification. In this stage, we verify *whether key physical phenomena occur in the correct order*. The correct physical sequence is an ordered series of events in a physical process that reflects causality, which represents the necessary prerequisites and temporal order of key physical phenomena. As an example, the egg should first touch the stone and then crack. Considering current models in *PhyGenBench* generally maintain outcome consistency (Huang et al., 2024) (e.g., the egg would not reassemble itself after it is broken). we approach this direction by investigating the order correctness from the keyframes (Figure 3 (b)), e.g., the keyframe of the egg hits the stone should be ahead of the keyframe of the broken egg.

Similar to the single image evaluation, we prompt GPT-4o to generate a retrieval prompt p_r and three physical-related questions (q_1, q_2, q_3). p_r is used to locate the keyframe (e.g., the moment the egg slightly touches the stone.). While q_1, q_2, and q_3 are questions to check the order correctness from the first frame to the keyframe, from the keyframe to the last frame, and from the first frame to the last frame, respectively. Similarly, we first use CLIPScore to locate the key frame I_i, then the order correctness scores of $\mathrm{S}_{\mathrm{before}}$ and $\mathrm{S}_{\mathrm{after}}$ are defined as:

$$\mathrm{S}_{\mathrm{before}} = \max_{i-2 \leq j \leq i} \left(\mathrm{VLM}(I_0, I_j, q_1) + \mathrm{VLM}(I_j, p_r) \right)$$
$$\mathrm{S}_{\mathrm{after}} = \max_{i \leq j \leq i+2} \left(\mathrm{VLM}(I_j, I_{-1}, q_2) + \mathrm{VLM}(I_j, p_r) \right)$$

q_3 assesses the overall physical sequence coherence of the video. The score of answering q_3 is defined as by $\mathrm{S}_{\mathrm{all}} = \mathrm{VLM}(I_0, I_{i-2:i+2}, I_{-1}, q_3)$, which evaluates the overall sequence (similar to the input video but using manually selected key frames). Here we employ GPT-4o or LLaVA-Interleave (Li et al., 2024) as the VLM-based evaluator $\mathrm{VLM}(\cdot)$, as they demonstrate exceptional multi-image comprehension capabilities. The overall score of whole physical order evaluation can be formulated as $\mathrm{S}_{\mathrm{order}} = \mathrm{S}_{\mathrm{before}} + \mathrm{S}_{\mathrm{after}} + \mathrm{S}_{\mathrm{all}}$

Overall Naturalness Evaluation. This stage aims to evaluate **the overall naturalness of the video**. we define naturalness as the dynamic progression that aligns with real-world physical phenomenons (Liao et al., 2024). For each prompt in *PhyGenBench*, we obtain a naturalness evaluation standard, denoted as g_{spec}, which is used to assess the naturalness for video. As shown in Figure 3 (c), we first refer to DEVIL (Liao et al., 2024) to establish a general evaluation standard: g_{gen}, applicable to all T2V prompts. Besides, we use each input T2V prompt p, the corresponding physical law l, and general evaluation standard g_{gen} to guide GPT-4o in generating a detailed evaluation standard: g_{spec}, for the given prompt. Finally, we require the VLM to score based on p, l, g_{spec}, and the corresponding video denoted by $I_{0:-1}$. Formally, we define the overall naturalness score as:

$$\mathrm{S}_{\mathrm{natural}} = \mathrm{VLM}(I_{0:-1}, p, l, g_{spec})$$

We implement the VLM-based evaluator $\mathrm{VLM}(\cdot)$ using InternVideo2 (Wang et al., 2024) and GPT-4o, both of which have demonstrated promising results in video understanding.

Table 1: **PCA correlation results with proposed *PhyGenEval* in video generation**. *PhyGenEval* is significantly closer to human feedback on *PhyGenBench* compared to other metrics.

Metric	Mechanics		Optics		Thermal		Material		Overall	
	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
DEVIL (Liao et al., 2024)	0.15	0.16	0.03	0.03	0.10	0.11	0.27	0.29	0.17	0.18
VideoPhy (Bansal et al., 2024)	0.00	−0.03	−0.15	−0.14	0.08	0.08	0.13	0.14	0.03	0.04
VideoScore (He et al., 2024b)	0.18	0.20	0.07	0.08	0.14	0.15	0.14	0.15	0.17	0.19
PhyGenEval	**0.72**	**0.75**	**0.76**	**0.77**	**0.73**	**0.75**	**0.81**	**0.84**	**0.78**	**0.81**

Overall Score. We first discretize $\mathrm{S}_{\mathrm{key}}$, $\mathrm{S}_{\mathrm{order}}$, and $\mathrm{S}_{\mathrm{natural}}$ from the three stages into a four-point scale, then take their average and apply floor rounding as the final score. For robust purposes, we evaluate $\mathrm{S}_{\mathrm{order}}$ with both GPT4o and LLaVA-Interleave and $\mathrm{S}_{\mathrm{natural}}$ with both GPT4o and InternVideo2. The final score is calculated as the ensemble of two methods. Detailed calculation protocols are provided in Appendix C.

5 Experiment

Experiments Setup. We evaluate 5 open-source models including OpenSora V1.2 (Zheng et al., 2024a), Lavie (Wang et al., 2023), CogVideoX 2b (Yang et al., 2024), CogVideoX 5b (Yang et al., 2024), and Vchitect2.0 (Wang et al., 2023), as well as proprietary models Kling (kli, 2024), Pika (Pik, 2023), and Gen-3 (gen, 2024). We compare our proposed metric with existing metrics or benchmarks: Videophy (Bansal et al., 2024), VideoScore (He et al., 2024b) and DEVIL (Liao et al., 2024) More Detailed information is provided in Appendix D.

For human evaluation, we compared the results across 8 T2V models. We randomly select 64 prompts from *PhyGenBench* and generate 64 videos for each T2V model. Therefore we need evaluation 512 videos. We ask three annotators to provide semantic and physical scores for each video[3]. Each annotator will give an integer score of 0-3 for the semantic and physical scores, and the final score is the average of the three scores and rounded up. Finally, we calculate the correlation between the human scores and automatic evaluation scores using Kendall's τ and Spearman's ρ. we put more detailed information about human evaluation in Appendix D.1.

Human Evaluation. As shown in Table 1, current video generation evaluation metrics largely overlook physical correctness. In contrast, *PhyGenEval* implements a detailed design for evaluating physical correctness, demonstrating strong correlations with human judgments across all categories. Its overall correlation coefficient reaches 0.81, indicating that *PhyGenEval* serves as an effective human-aligned physical commonsense correctness evaluator for *PhyGenBench*. We put more results in Appendix D.2

We conduct several case studies to illustrate the differences between various metrics more clearly. As shown in Figure 4, (a) and (f) reveal that VideoScore and DEVIL are prone to misclassifying videos that have smooth and consistent motion but violate fundamental physical laws. Specifically, as for (a), when *"an egg exhibits rubber-like elasticity upon impact with a rock instead of breaking,"* these metrics incorrectly evaluate it as physically correct. VideoPhy exhibits similar limitations. In (c), it incorrectly assesses *"a rock floating on water instead of sinking"* as physically correct. Furthermore, our analysis reveals a major flaw in these three methodologies: they cannot incorporate domain-specific physical commonsense. As illustrated in (e), where *"the flame from burning copper appears red instead of green,"* these metrics fail to identify the mistake. This indicates their inability to incorporate domain-specific physical commonsense. In contrast, *PhyGenEval* demonstrates a robust integration of physical commonsense and comprehensive video content analysis, resulting in more accurate and physically consistent evaluations in *PhyGenBench*.

Quantitative Evaluation. We conduct extensive experiments on a wide range of popular video generation models. As illustrated in Table 2, even the best-performing model, Gen-3, only attains a PCA score of 0.51 on *PhyGenBench*. This indicates that even for prompts containing obvious

[3]Note that we ask the annotators to focus on the correctness of the physical phenomena for physical scores.

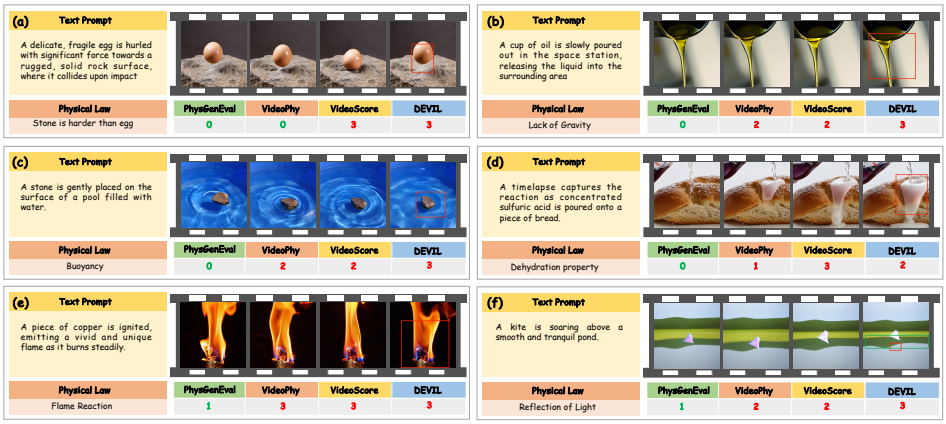


Figure 4: Different video generation evaluation metric in *PhyGenBench*. Except for the proposed *PhyGenEval*, the current methods cannot reasonably assess the correctness of physical commonsense in videos from *PhyGenBench*.

Table 2: **Evaluation results of PCA with the proposed *PhyGenEval* in videos generated by several models** . The results reveal that all models score very low in physical commonsense accuracy. The scores are normalized to a range of 0-1.

Model	Size	Mechanics(↑)	Optics(↑)	Thermal(↑)	Material(↑)	Average(↑)	Human(↑)
CogVideoX (Yang et al., 2024)	2B	0.38	0.43	0.34	0.39	0.39	0.31
CogVideoX (Yang et al., 2024)	5B	0.39	0.55	0.40	0.42	0.45	0.37
Open-Sora V1.2 (Zheng et al., 2024a)	1.1B	0.43	0.50	0.44	0.37	0.44	0.35
Lavie (Wang et al., 2023)	860M	0.30	0.44	0.38	0.32	0.36	0.30
Vchitect 2.0 (Wang et al., 2023)	2B	0.41	0.56	0.44	0.37	0.45	0.36
Pika (Pik, 2023)	-	0.35	0.56	0.43	0.39	0.44	0.36
Gen-3 (gen, 2024)	-	**0.45**	0.57	0.49	**0.51**	**0.51**	**0.48**
Kling (kli, 2024)	-	**0.45**	**0.58**	**0.50**	0.40	0.49	0.44

physical commonsense, current T2V models struggle to generate videos that comply with intuitive physics. It indirectly reflects that these models are still far from achieving the world simulator.

Furthermore, we identify the following key observations: **1):** Across various categories of physical commonsense, all models consistently demonstrate superior performance in the domain of optics compared to other areas. Notably, Vchitect2.0 and CogVideoX-5b achieve a PCA score in the optics domain comparable to that of closed-source models. We posit that this superior performance in the optics domain can be attributed to the abundant and explicit representation of optical knowledge in pre-training datasets, thereby enhancing the model's comprehension in this area. **2):** Kling and Gen-3 exhibit significantly higher performance compared to other models. Specifically, Gen-3 demonstrates a robust understanding of material properties, achieving a score of 0.51, which substantially surpasses other models. Kling performs particularly well in thermal, attaining the highest score of 0.50 in this domain. **3):** Among open-source models, Vchitect2.0 and CogVideoX 5b perform comparatively well, both exceeding the performance level of Pika. In contrast, Lavie consistently exhibits lower physical correctness across all categories.

Qualitative Evaluation. The different video cases for 4 physical commonsense categories are illustrated in Figure 5. Our main observations are as follows: In mechanics, the models struggle to generate simple physically accurate phenomenons. As shown in Figure 5, all models fail to depict the glass ball sinking in water. As for (b), instead showing it floating on the surface, OpenSora and Gen-3 even produce videos where the ball is suspended. Additionally, the models do not capture special physical phenomenonss, such as the state of water in zero gravity, as seen in (a). In optics, the models perform relatively better. (c) and (d) show the models handling reflections of balloons in water and colorful bubbles, though OpenSora and CogVideoX still produce reflections with noticeable distortions in (d). In thermal, the models fail to generate accurate videos of phase transitions. For the melting phenomenon in (e), most models show incorrect results, with CogVideoX even producing

Figure 5: Qualitative comparisons of four categories. Current models perform relatively well in generating optical phenomenons but are weaker in mechanics, thermal, and material properties.

a video where the ice cream increases in size. Similar errors appear in the sublimation process in (f), with only Gen-3 showing partial understanding. Regarding material properties, (g) shows all models failing to recognize that an egg should break when hitting a rock, with Kling displaying the egg bouncing like a rubber ball. For simple chemical reactions, such as the black bread experiment in (h), none of the models demonstrate an accurate understanding of the expected reaction.

Ablation Study. We conduct a detailed robustness analysis of the design elements in PhyGenEval, including the role of each level in the three-tier evaluation framework and the impact of the two-stage strategy proposed in overall naturalness evaluation. Experimental results show that the key designs of *PhyGenEval* are essential. Detailed results are provided in Appendix D.3.

6 Discussion

To explore potential solutions for the challenges posed by *PhyGenBench*, We focus on widely used and proven-effective methods such as scaling laws (Kaplan et al., 2020), prompt engineering (Fu et al., 2024), and some method like Venhancer (He et al., 2024a) aimed to improve general video quality (Huang et al., 2024). Through quantitative and qualitative analysis, we find: 1) Scaling up models can solve some issues but still fails to handle dynamic physical phenomenons, which we believe requires extensive training on synthetic data. 2) Prompt engineering like (Fu et al., 2024) only solves a few simple issues (e.g., flame color), highlighting the difficulty and significance of *PhyGenBench*. 3) While some methods improve general video quality, they do not enhance the model's understanding of physical commonsense. More detailed results are in Appendix E.

7 Conclusion

In this paper, we explore the gap between current T2V models' understanding of physical commonsense and their role as world simulators. To achieve this, we introduce *PhyGenBench* and *PhyGenEval*. *PhyGenBench* is a benchmark specifically designed to assess models' understanding of physical commonsense, featuring various physical laws and simple, clear physical phenomenons. Alongside *PhyGenBench*, we propose a novel three-tier hierarchical evaluation framework called *PhyGenEval* to automate the evaluation process. Experimental and analytical results show that current T2V models struggle to generate videos that align with physical commonsense, highlighting a significant gap from world simulation. Moreover, simply scaling up models or applying prompt engineering fails to address issues in *PhyGenBench*, such as those involving dynamics.

REFERENCES

Pika, 2023. URL https://www.pika.art/.

Gen-3, 2024. URL https://runwayml.com/blog/introducing-gen-3-alpha/.

Kling, 2024. URL https://kling.kuaishou.com/.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

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

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

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.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.

David Halliday, Robert Resnick, and Jearl Walker. *Fundamentals of physics*. John Wiley & Sons, 2013.

Ahmad Harjono, Gunawan Gunawan, Rabiatul Adawiyah, and Lovy Herayanti. An interactive e-book for physics to improve students' conceptual mastery. *International Journal of Emerging Technologies in Learning (iJET)*, 15(5):40–49, 2020.

Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024a.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L. Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14963–14973, June 2023.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *arXiv preprint arXiv:2407.01094*, 2024.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision ECCV*, 2024b.

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

Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, and Sai Rajeswar. Multimodal foundation world models for generalist embodied agents. *arXiv preprint arXiv:2406.18043*, 2024.

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.

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.

K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

Norman Swartz. *The concept of physical law*. Cambridge University Press, 1985.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

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.

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.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

Justin N Wood, Tomer D Ullman, Brian W Wood, Elizabeth S Spelke, and Samantha MW Wood. Object permanence in newborn chicks is robust against opposing evidence. *arXiv preprint arXiv:2402.14641*, 2024.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

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

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *arXiv preprint arXiv:2406.18522*, 2024.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024a. URL `https://github.com/hpcaitech/Open-Sora`.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024b.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.

A Related work

A.1 Benchmarks For Physical Understanding

Recent benchmarks such as SuperCLEVR-Physics Li et al. (2023), ContPhy Zheng et al. (2024b), and Physion Bear et al. (2021) have significantly advanced the evaluation of AI's physical reasoning, focusing on understanding and predictive tasks. SuperCLEVR-Physics emphasizes reasoning about dynamic properties like velocity and collisions in 4D scenes, ContPhy expands the evaluation scope to include diverse physical properties, such as mass and density, within continuum settings. It underscores the limitations of existing AI models in handling soft-body dynamics. And Physion evaluates models' ability to predict physical phenomena like collisions and motion while benchmarking against human behavior. However, these works primarily target understanding or prediction rather than generative capabilities. In contrast, our work introduces PhyGenBench, a comprehensive benchmark designed to evaluate whether Text-to-Video (T2V) models can generate videos that adhere to physical commonsense. Unlike existing benchmarks, our work highlights the generative challenges in intuitive physics, revealing critical gaps in current T2V models and underscoring the need to advance physical commonsense for applications beyond entertainment.

B PhyGenBench Details

B.1 Detailed Overview

A fine-grained analysis of the dataset is essential for a comprehensive understanding of the benchmark. As shown in Table 3, *PhyGenBench* covers 4 major domains in physics, encompassing 27 representative physical laws, which enables it to provide a more comprehensive and fine-grained evaluation of models' physical capabilities. We generated 1280 videos by evaluating 8 advanced models. Additionally, our captions encompass totally 165 unique objects and 42 unique actions with an average length of 18.75 words.

Table 3: Details of *PhyGenBench*

Statistic	**Number**
Physical Laws	27
Domains	4
Optics	50
Mechanics	40
Thermal	30
Material Properties	40
Total Captions	160
Total T2V Models	8
Total Generated Videos	1280
Unique Objects	165
Unique Actions	42
Average Length of Caption	18.75

B.2 Difference between Videophy and Ours

VIDEOPHY Bansal et al. (2024) comprises 688 curated simple prompts that focus on interactions between three types of physical materials: solid-solid, solid-fluid, and fluid-fluid, but lack annotations of physical laws. The dataset is designed to evaluate a model's understanding of physical commonsense, featuring a limited range of physical phenomenons such as rigid body interactions, fluid dynamics, and contact forces. We are better suited than Videophy for evaluating physical commonsense due to two significant differences.

First As shown in Figure 2, *PhyGenBench* includes 160 carefully crafted prompts across 27 distinct physical laws, spanning four fundamental domains, which comprehensively assess a model's understanding of physical commonsense. While Videophy primarily focuses on interactions between solid-fluid, solid-solid, and fluid-fluid, limiting its coverage and overlooking common physical laws such as phase transitions and basic material properties. What' more, Videophy lacks annotations of physical laws making it hard for VLM model to evaluate. Second, as shown in Table 4, the average SA score of *PhyGenBench* (0.80) significantly outperforms that of Videophy (0.63). This indicates that *PhyGenBench* prompts are well-suited and easy for T2V models to generate high-quality, well-aligned videos, which benefits evaluation of physical correctness. In contrast, as shown in Figure 6, We find that prompts from Videophy pose challenges for T2V models in generating text-aligned and high-quality videos for two main reasons: 1. The prompts lack detail and specificity. For instance, *"A tissue blots a tear from an eye"* is overly simplistic (without augmentation). Modern T2V

models, such as CogVideo5B Yang et al. (2024), are typically trained with longer and more descriptive captions, which enhance their ability to comprehend and generate content based on prompts. 2. The scenes are often complex and unrealistic. For example, "The wristwatch knob winds the inner spring tightly" describes a process involving intricate internal mechanisms that are not visible externally. As a result, it is exceedingly difficult for T2V models to generate such scenes accurately.

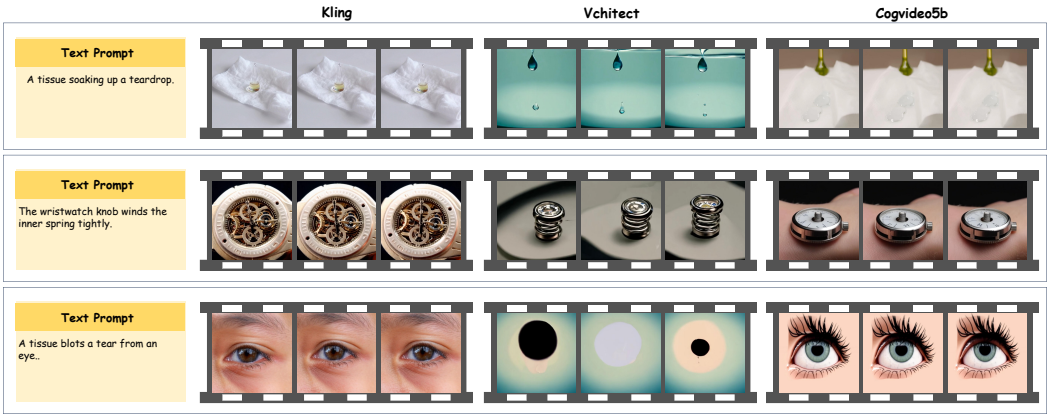


Figure 6: **Samples of videos generated by Kling, Vchitect, and Cogvideo5b in Videophy.** All T2V models struggle to achieve proper text alignment and produce high-quality videos, making it meaningless to evaluate physical correctness in Videophy.

Table 4: **Comparison of SA results for video generation between Videophy and *PhyGenBench*.** We randomly select 64 prompts from both Videophy and *PhyGenBench*, use different T2V models to generate videos, and then ask annotators to score based on our cretiera in Figure 10. The results show that *PhyGenBench* 's SA scores significantly outperform Videophy.

Model	Size	Videophy(↑)	*PhyGenBench* (↑)
CogVideoX (Yang et al., 2024)	5B	0.48	0.78
Vchitect 2.0	2B	0.63	0.84
Kling	-	**0.77**	**0.89**
Average	-	0.63	0.80

C PhyGenEval Details

C.1 Semantic alignment details

To reduce the complexity for VLM models to evaluate sementic correctness of generated videos between prompts, we adopt a two-stage strategy. Initially, we employ GPT-4o to extract objects and actions from the original text prompt. Subsequently, we employ GPT-4o to determine whether the extracted objects are present in the video and to verify the occurrence of specified actions. For each video, GPT-4o first assesses the presence of the objects mentioned in the prompt (e.g., *"egg"*) within the video frames. This evaluation is performed according to Question 1 (Q1), where GPT-4o assigns a score from 0 to 2 based on the completeness of object presence: a score of 2 is given if all the objects are present, 1 if some of the objects are missing, and 0 if none of the objects appear in the video. After determining object presence, GPT-4o moves on to Question 2 (Q2) to check if the specified action (e.g., *"pour out"*) is performed in the video. It assigns a binary score (0 or 1) depending on whether the action is present (1) or absent (0). Finally, these scores are combined to form the overall semantic alignment score. we put more details about other metric baselines in Appendix D.1.

C.2 PHYSICAL COMMONSENSE ALIGNMENT DETAILS

In this section, we use the same notation as in Section 4.2 and provide a more detailed description of the calculation and design of the method.

Key Phenomena Detection. We categorize the T2V prompts into monotonic processes (eg. *"melting with increasing temperature"*) and non-monotonic processes (eg. *"an egg hitting a rock"*) based on the physical phenomena they represent. For prompt with monotonic processes, we only consider using the *"Last Frame"* as the retrieval prompt, resulting in a single question. We can directly calculate $\mathrm{VLM}(\mathrm{Img}_j, Q)$, where the score for the corresponding video of this prompt ranges from 0 to 1. For prompt with non-monotonic processes, we consider both the intermediate key frames and the Last Frame, resulting in two questions. For the intermediate key frames, we calculate $\mathrm{VLM}(\mathrm{Img}_j, Q) + \mathrm{VLM}(\mathrm{Img}_j, P_r)$, which ranges from 0-2. Consequently, the score range for videos corresponding to this prompt is 0 to 3.

For specific calculatation, we need to calculate $\mathrm{VLM}(\mathrm{I}_j, p_r)$ and $\mathrm{VLM}(\mathrm{I}_j, q)$, where Img_j is the j-th frame in the video. For $\mathrm{VLM}(\mathrm{I}_j, p_r)$, the calculation involves assessing the matching degree between the key frame and the retrieval prompt, which can be directly obtained using the original calculation method in (Lin et al., 2024). For $\mathrm{VLM}(\mathrm{I}_j, q)$, we follow the computation approach from ChronoMagicBench (Yuan et al., 2024), we derive $\mathrm{VLM}(\mathrm{I}_j, q)$ by determining the ratio of the VQAScore for the affirmative statement to the combined VQAScores for both the affirmative and negative statements. We perform the calculations of $\mathrm{VLM}(\mathrm{I}_j, p_r)$ and $\mathrm{VLM}(\mathrm{I}_j, q)$ for each key frame within the specified range to obtain the physical correctness score for the problem.

Key Sequence Verification. For this stage, which we've primarily introduced in Section 4, we focus on key calculation points. The score calculation formula for q_1 is $\mathrm{S}_{\mathrm{before}} = \max_{i-2 \le j \le i} \left(\mathrm{VLM}(\mathrm{I}_0, \mathrm{I}_j, q_1) + \mathrm{VLM}(\mathrm{I}_j, p_r)\right)$. Here, $\mathrm{VLM}(\mathrm{I}_j, p_r)$ determines if the retrieved key frame satisfies the retrieval prompt,as the physical phenomenon should occur in the keyframe primarily located in Key Phenomena Detection, which is crucial for Key Sequence Verification (e.g.the expected physical phenomenon of egg cracking should occur in the keyframe when the egg hits the stone, rather than other frames when the egg is in the air or else). $\mathrm{VLM}(\mathrm{I}_0, \mathrm{I}_j, q_1)$ assesses the correctness of the Key Sequence order in the video. Notably, we calculate $\mathrm{VLM}(\mathrm{I}_j, p_r)$ using VQAScore, yielding a decimal between 0 and 1, while $\mathrm{VLM}(\mathrm{I}_0, \mathrm{I}_j, q_1)$ employs VLM (GPT-4V or LLaVA-Interleave) for question-answering, scoring 1 or 0 based on the model's Yes or No response.

Overall Naturalness Evaluation. Here we mainly explain how to get the score of this part based on the evaluation results under the two-stage strategy described in Section 4. Specifically, we ask the video-based VLM to select the most appropriate option for the video according to the detailed scoring criteria generated by the LLM, and then we map the options to scores (Completely Fantastical to Almost Realistic corresponds to 0-3 points)

Overall Score. We detail the discretization and calculation process of the scores here. In the stage of key phenomena detection, we categorize the prompts into monotonic and non-monotonic processes based on the physical phenomena they represent. For monotonic processes, the score range is 0-1, which we directly discretize by averaging into integer values from 0-3. Specifically, for non-monotonic processes with a score range of 0-3, we discretize the scores to $[1, 1.5, 2.25]$. This is because no points should be awarded if the physical phenomena are incorrect ($\mathrm{VLM}(\mathrm{I}_j, p_r) = 1$ and $\mathrm{VLM}(\mathrm{I}_j, q) = 0$), even with accurate retrieval. (e.g., The egg hits the stone and does not break)

In the stage of key sequence verification, we have three multi-image problems. One point is awarded for each correct answer, resulting in a final integer score from 0-3. Similar to the stage, of key phenomena detection we need to consider both the accuracy of key frame retrieval and the physical question answering. Therefore, we design the following: for Q_1, when $\max_{i-2 \le j \le i} \left(\mathrm{VLM}(\mathrm{I}_0, \mathrm{I}_j, q_1) + \mathrm{VLM}(\mathrm{I}_j, p_r)\right)$ and $\mathrm{VLM}(\mathrm{I}_j, p_r) > 0.5$, the question is considered correct. The process for q_2 is similar. For q_3, it is marked correct when $\mathrm{VLM}(\mathrm{I}_0, \mathrm{I}_{i-2:i+2}, \mathrm{I}_{-1}, q_3)$.

In the stage of overall naturalness evaluation, as we require video-based direct option selection, choosing Completely Fantastical, Clearly Unrealistic, Slightly Unrealistic, and Almost Realistic is scored as 0, 1, 2, and 3 points respectively. Finally, we average all scores and round down to obtain the final score.

Table 5: Details about evaluation models. The table shows duration, FPS, and resolution for each model.

Model	Duration (s)	FPS	Resolution
Open-Sora 1.2 (Zheng et al., 2024a)	4	24	1280 × 720
CogVideoX 2b	6	8	720 × 480
CogVideoX 5b	6	8	640 × 360
Lavie	4	8	512 × 320
Vchitect2.0	5	8	768 × 432
Pika (Pik, 2023)	3	24	1280 × 720
Gen-3 (gen, 2024)	11	24	1280 × 768
Kling (kli, 2024)	5	30	1280 × 720

For the ensamble operation, in order to reduce the bias caused by using a single VLM at a certain stage, we ensemble the results of PhyGenEval using open source models or closed source models. Specifically, we average the two results and round them down.

D EXPERIMENT

D.1 EXPERIMENTS SETUP

T2V model Implementation details. Open-Sora 1.2 (Zheng et al., 2024a) is an open-source project with the goal of reproducing Sora. CogVideoX 2b Yang et al. (2024) and CogVideoX 5b are large-scale diffusion transformer models for text-to-video generation, incorporating a 3D Variational Autoencoder (VAE) for efficient video compression and an expert transformer with Expert Adaptive LayerNorm to improve text-video alignment. LaVie Wang et al. (2023) is a cascaded video latent diffusion model. Vchitect2.0 Wang et al. (2023), developed by the Shanghai AI Lab, is an advanced video generation model featuring a Parallel Transformer architecture to scale up video diffusion models and empower video creation.

Evaluation Metrics details. We compare our proposed *PhyGenEval* with some evaluation metrics from previous methods like VideoPhy (Bansal et al., 2024) and VideoScore (He et al., 2024b). VideoPhy fine-tunes a VLM with the VIDEOPHY dataset proposed by themselves, which includes human feed back about the semantic alignment and dynamic motion correctness about videos. VideoScore is trained on the VIDEOFEEDBACK dataset proposed by themselves, Initialized from the Mantis model. VideoScore provides automatic assessments of video quality based on human scoring criteria. To compare with *PhyGenEval* on SA and PCA, We only choose the text alignment and fact consistency criteria. Specifically, for the semantic alignment evaluation, we compare the Grid-LLaVA method proposed by T2V-CompBench, which extends the LLaVA (Liu et al., 2024a) model to handle multi-frame inputs by sampling 6 frames uniformly from a video to create an image grid. For the physical commonsense alignment evaluation, we also compare with DEVIL (Liao et al., 2024), which uses Gemini 1.5 Pro (Reid et al., 2024) to assess the overall naturalness of videos and applies the same scoring standard prompt to all videos.

Furthermore, to evaluate the effectiveness of our *PhyGenEval* designs, we conduct a large amount of ablation studies and pue more details in Appendix D.3.

Human evaluation details. Here, we provide a detailed explanation of the human evaluation described in Section 5. Specifically, we require annotators to score based on the standards outlined in Figure 10, covering both semantic alignment and physical commonsense alignment. For example, as for the video shown in Figure 10, The egg bounces off the rock like a rubber ball, completely violating physical laws like dynamics, the annotator gives a score of 0 for physical commonsense alignment. However, since the video fully includes the egg, the rock, and the collision action, the annotator gives a score of 3 for semantic alignment.

Table 6: **SA correlation results with proposed *PhyGenEval* in video generation**. A higher score indicates better performance for a category. **Bold** stands for the best score,

Metric	Mechanics		Optics		Thermal		Material		Overall	
	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
VideoPhy (Bansal et al., 2024)	0.20	0.25	0.03	0.03	0.20	0.24	0.18	0.22	0.13	0.17
VideoScore (He et al., 2024b)	0.14	0.16	−0.13	−0.14	0.23	0.02	0.02	0.02	0.05	0.05
Grid-LLaVA (Sun et al., 2024)	0.39	0.43	0.45	0.49	0.30	0.33	0.22	0.26	0.35	0.39
PhyGenEval (Grid-LLaVA)	0.35	0.38	0.46	0.48	0.41	0.44	0.42	0.45	0.42	0.44
PhyGenEval	0.48	0.52	0.64	0.67	0.46	0.49	0.47	0.50	0.53	0.56

Table 7: **SA evaluation results with proposed *PhyGenEval* in video generation**. Both machine and human evaluations indicate that most models achieve good semantic scores on *PhyGenBench*. This suggests that the scenarios in *PhyGenBench* are simple enough to clearly reflect physical phenomena.

Model	Size	Mechanics(↑)	Optics(↑)	Thermal(↑)	Material(↑)	Average(↑)	Human(↑)
CogVideoX (Yang et al., 2024)	2B	0.63	0.67	0.61	0.63	0.64	0.64
CogVideoX (Yang et al., 2024)	5B	0.78	0.88	0.78	0.64	0.78	0.78
Open-Sora V1.2 (Zheng et al., 2024a)	1.1B	0.73	0.85	0.82	0.73	0.79	0.70
Lavie (Wang et al., 2023)	860M	0.47	0.63	0.73	0.53	0.58	0.55
Vchitect 2.0 (Wang et al., 2023)	2B	**0.92**	0.89	0.77	0.74	0.84	0.84
Pika (Pik, 2023)	-	0.63	0.81	0.73	0.69	0.72	0.65
Gen-3 (gen, 2024)	-	0.84	**0.93**	0.82	**0.78**	**0.85**	0.86
Kling (kli, 2024)	-	0.88	0.91	**0.87**	0.74	**0.85**	**0.89**

D.2 Quantitative Evaluation

Comparison result about semantic alignment. Here we design a new baseline *PhyGenEval* (Grid-LLaVA) to illustrate the superiority of the method, which uses the two-stage strategy proposed in *PhyGenEval* from Appendix C.1, but replaces the VLM with Grid-LLaVA proposed in T2V-CompBench (Sun et al., 2024). As shown in Table 6, *PhyGenEval* achieves the highest correlation scores across all categories, demonstrating its effectiveness as a human-aligned semantic commonsense correctness evaluator for *PhyGenBench*. Compared to other methods, *PhyGenEval* consistently outperforms previous baselines like VideoPhy, VideoScore, and Grid-LLaVA. Specifically, *PhyGenEval* obtains an overall Kendall's τ of 0.53 and a Spearman's ρ of 0.56, surpassing the Grid-LLaVA (τ: 0.35, ρ: 0.39). The results clearly show that our *PhyGenEval* design provides a more accurate and reliable semantic commonsense evaluation in *PhyGenBench*.

Quantitative result about semantic alignment. As shown in Table 7 , nearly all models achieve relatively high SA scores, whether evaluated by machines or humans. This suggests that the scenarios in *PhyGenBench* are relatively straightforward, making it easier to assess physical commonsense. Among all the models, Kling achieved the highest SA score, with a human evaluation score of 0.89, reflecting its strong instruction understanding and video generation capabilities.

D.3 Ablation study

The Component in *PhyGenEval* on physical commonsense alignment evaluation. We conduct a series of ablation studies to demonstrate the necessity of our method design by examining its correlation with human evaluation results, similar to those described in Section 5. Specifically, we compare: 1) The effectiveness of two-stage evaluation method proposed in Section 4.2 2) The effect of the various stages of *PhyGenEval*, as proposed in Section 4.2; 3) Performance differences when using various VLMs and their ensembles in *PhyGenEval*, as outlined in Section 4.2. 4) The larger open models in PhyGenEval. Notice that *PhyGenEval* for physical commnonsense alignment evaluation consists of three stages: key phenomena Detection, key sequence verification, and overall naturalness evaluation. And We denote them as *PhyGenEval*-S, *PhyGenEval*-M, and *PhyGenEval*-V based on the VLM they used.

1) We demonstrate that employing a two-stage strategy, as outlined in Section 4.2, yields superior results when assessing the physical commonsense correctness of the entire video compared to one-

stage strategy. Specifically, the one-stage strategy refers to not using LLM to rewrite the scoring template, but instead applying a single scoring template for all prompts' corresponding videos, allowing the VLM to score them. This method is proposed in DEVIL (Liao et al., 2024). To verify the superiority of the two-stage strategy, we use InternVideo2 and GPT-4o as VLMs and perform both the one-stage and two-stage strategies. We label these as *PhyGenEval*-V(Intern) and *PhyGenEval*-V(GPT-4o), respectively. As shown in Table 8, the evaluation results produced by the two-stage strategy are more consistent with human judgments for both InternVideo2 and GPT-4o. We attribute this improvement to the incorporation of LLM (GPT-4o) for better comprehension of physical commonsense text, which effectively reduces the complexity of the task for VLMs in evaluating the physical correctness of videos.

2) *PhyGenEval* for physical commnonsense alignment evaluation consists of three stages. We investigate the contribution of each stage to the final performance. Table 9 presents results using one or two stages (employing ensemble strategies when multiple VLMs are applicable). We find that optimal performance is achieved only when all three stages are used concurrently, demonstrating the rationale behind *PhyGenEval*'s design.

3) Given potential biases in single models and the costs associated with closed-source models, we offer two *PhyGenEval* computation methods: using GPT-4o or alternative open-source models (LLaVA-Interleave (Li et al., 2024) and InternVideo2 (Wang et al., 2024)). Table 10 shows that even using only small scale open-source models achieves a high correlation coefficient of 0.66. Notably, ensembling both methods yields the best results. Considering *PhyGenBench*'s relatively small size, we find this computational cost acceptable. Therefore we recommend users ensemble these methods.

4) We explore the performance of larger open-source models. Specifically, we replace LLaVA-Interleave used in the Physics Order Verification stage of PhyGenEval (Open-S) with InternVL-Pro (78B), denoted as PhyGenEval (Open-L). Additionally, we ensemble PhyGenEval (Open-L) with PhyGenEval (Open-S), denoted as PhyGenEval (Open-Ensemble). Results show that compared to smaller open-source models, the overall alignment coefficient improves from 0.66 to 0.72, indicating that the method remains reproducible even when using exclusively open-source models. We believe that as open-source models continue to advance, they can achieve even better performance within PhyGenEval.

The Component in *PhyGenEval* on semantic alignment evaluation. we also perform necessary ablation experiments to validate the necessity of our SA evaluation design. Specifically, we compare: 1) VLM Model Selection: We leverage GPT-4o (Achiam et al., 2023) as a more robust VLM model for SA evaluation. 2) Effectiveness of our two-stage evaluation method proposed in Appendix C.1

1) As shown in Table 6, using GPT-4o in *PhyGenEval* is much better than using LLaVA, which achieve a higher Kendall's τ of 0.53 compared to 0.42, and a higher Spearman's ρ of 0.56 versus 0.44. This indicates a stronger alignment between GPT-4o's evaluations and human annotations compared to open-source vlm models like Grid-LLaVA (Sun et al., 2024), justifying its selection as the preferred VLM model in the SA evaluation design. Since *PhyGenBench* includes a limited number of prompts, we believe that the cost of using GPT-4o is acceptable relative to the improvement in performance.

2) To validate the effectiveness of the two-stage strategy, we compare it with the method in T2V-CompBench (Sun et al., 2024), which directly uses Grid-LLaVA to apply the same scoring standard prompt for semantic alignment evaluation across all videos. For fairness, we also use Grid-LLaVA but implement the two-stage strategy proposed in Appendix C.1. As shown in Table 6, *PhyGenEval*-Grid-LLaVA outperforms Grid-LLaVA, achieving a higher Kendall's τ score of 0.42 compared to 0.35, and a higher Spearman's ρ score of 0.44 versus 0.39. This result demonstrates the effectiveness of our Two Stage Evaluation Method. By decomposing the evaluation into object detection and action detection, we effectively reduces the complexity of the task for VLMs in evaluating the sementic correctness of videos.

E DISCUSSION

Table 8: Comparison of PCA correlation results of the two-stage strategy for the video stage in *PhyGenEval*

Metric	Mechanics		Optics		Thermal		Material		Overall	
	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
One Stage Strategy										
PhyGenEval-V(Intern)	−0.03	−0.04	−0.20	−0.21	−0.26	−0.27	0.06	0.06	−0.10	−0.11
PhyGenEval-V(GPT)	0.39	0.41	0.11	0.12	0.19	0.20	0.36	0.39	0.19	0.21
Two Stage Strategy										
PhyGenEval-V(Intern)	0.01	0.01	0.06	0.06	0.08	0.08	0.10	0.11	0.07	0.08
PhyGenEval-V(GPT)	0.47	0.51	0.50	0.53	0.46	0.49	0.53	0.58	0.53	0.58

Table 9: Comparison of PCA correlation results using each stage in *PhyGenEval*

Metric	Mechanics		Optics		Thermal		Material		Overall	
	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
PhyGenEval-S	0.50	0.54	0.43	0.45	0.50	0.54	0.72	0.77	0.56	0.61
PhyGenEval-M	0.46	0.49	0.49	0.53	0.55	0.59	0.53	0.57	0.55	0.60
PhyGenEval-V	0.26	0.30	0.44	0.47	0.33	0.35	0.48	0.52	0.42	0.46
PhyGenEval-SM	0.58	0.61	0.47	0.50	0.58	0.62	0.66	0.70	0.60	0.64
PhyGenEval-SV	0.56	0.59	0.41	0.43	0.58	0.60	0.70	0.74	0.59	0.62
PhyGenEval-MV	0.50	0.53	0.50	0.53	0.53	0.57	0.60	0.64	0.57	0.61
PhyGenEval	**0.72**	**0.75**	**0.76**	**0.77**	**0.73**	**0.75**	**0.81**	**0.84**	**0.78**	**0.81**

Table 10: Comparison of PCA correlation results using different models such as GPT-4o or open-sourced models in *PhyGenEval*

Metric	Mechanics		Optics		Thermal		Material		Overall	
	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
PhyGenEval (Open-S)	0.54	0.57	0.59	0.62	0.55	0.58	0.65	0.69	0.62	0.66
PhyGenEval (Open-L)	0.56	0.59	0.61	0.63	0.59	0.61	0.67	0.71	0.65	0.69
PhyGenEval (Open-Ensemble)	0.58	0.62	0.63	0.65	0.62	0.64	0.70	0.73	0.67	0.72
PhyGenEval (GPT4o)	0.59	0.63	0.53	0.57	0.64	0.68	0.73	0.77	0.66	0.71
PhyGenEval (Ensemble)	**0.72**	**0.75**	**0.76**	**0.77**	**0.73**	**0.75**	**0.81**	**0.84**	**0.78**	**0.81**

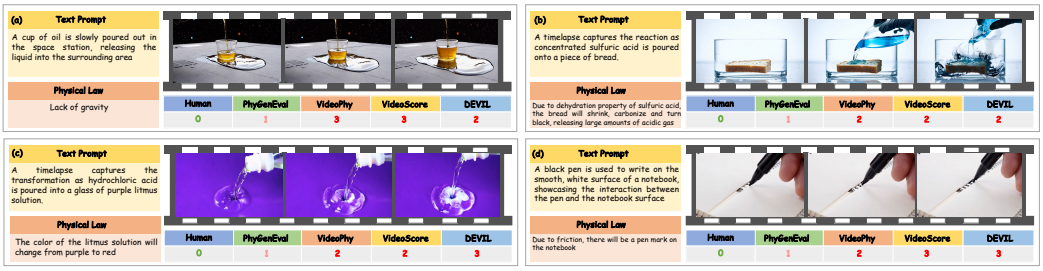


Figure 7: Visualization of some PhyGenEval error cases.

Error case analysis. As shown in Figure 7, we visualize some error cases where both PhyGenEval and competing methods like DEVIL fail to correctly identify the physical realism of the videos. These error cases are often caused by **confusing but iconic physical phenomena** in the videos that do not align with the correct progression of physical processes (e.g., in the erroneous case of the "burnt bread" experiment, black coloration appears but does not align with the expected phenomenon), leading to misjudgments. However, even in these cases, PhyGenEval remains closer to human ratings compared to other methods.

The Impact of Scaling on Physical Commonsense in Video Generation. Scaling laws have been extensively validated in video generation models (Kaplan et al., 2020). We investigate their efficacy in addressing the challenges of physical commonsense presented in *PhyGenBench*. As shown in Table 2, CogVideo 5B demonstrates improvements over CogVideo 2B, albeit with limited progress in the Mechanics category. Our qualitative analysis, illustrated in Figure 8, reveals significant advancements in static scenes with CogVideo 5B. It accurately captures complex phenomena such as colorful bubbles resulting from interference and diffraction, and oxidation-induced rusting of iron. In thermal, despite imperfections, CogVideo 5B generates more realistic boiling simulations compared to its predecessor. However, both models struggle with simple motion dynamics, exemplified by their inability to accurately depict a bouncing football. We posit that while scaling up enhances the model's capacity to generate videos that align with physical commonsense for individual objects, it may be insufficient for physical phenomenons involving dynamic physical laws. Addressing these challenges likely requires extensive training on carefully curated synthetic data, as suggested by (Liu et al., 2024b). This approach could potentially bridge the gap in the model's grasp of fundamental physical laws.

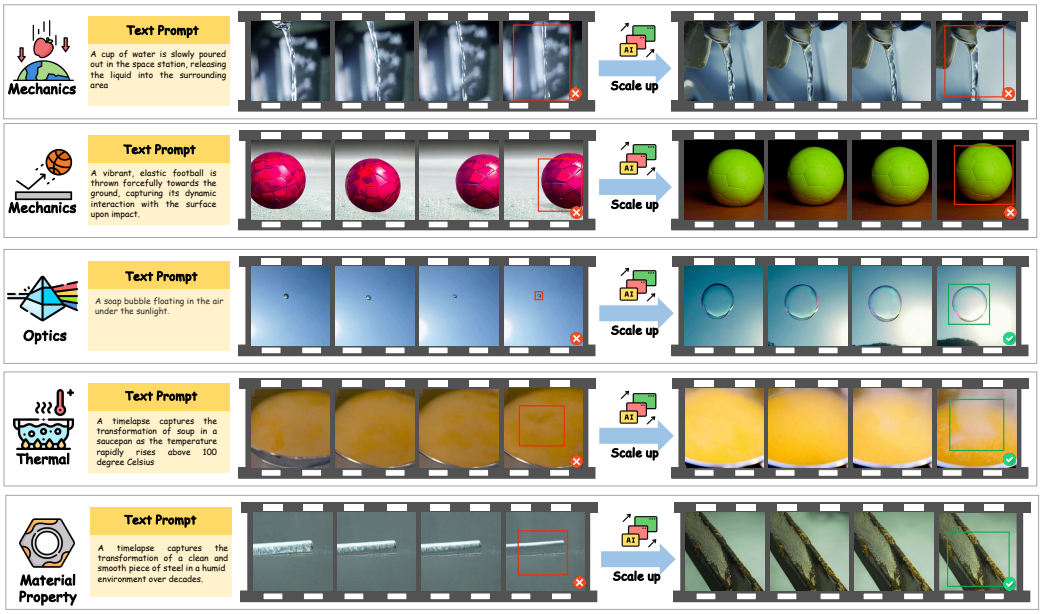


Figure 8: The qualitative comparison of CogVideoX 2B and CogVideoX 5B. The result shows that simply scaling up can solve some issues, but dynamic physical phenomenons involving the design of motion patterns remain challenging.

Rewriting prompt. We aim to explore whether GPT-augmented prompts can address the *PhyGenBench* challenges. Specifically, we rewrite the original prompts using GPT, adding expected physical outcomes and processes. For example, after *"A bottle of juice is slowly poured out in the space station, releasing the liquid into the surrounding area"*, we add *"The liquid forms floating globules, spreading out and moving randomly through the air."* in the end.

As shown in Table 11, we use CogVideoX 5b and Kling as representative models for open-source and closed-source systems, respectively, to conduct tests. The results indicate that prompt rewriting

Table 11: **Evaluation results of PCA using the proposed *PhyGenEval* after rewriting prompts** . The results indicate that although using rewritten prompts leads to some improvement, it is still insufficient to address the challenges highlighted by *PhyGenBench*.

Model	Size	Mechanics(↑)	Optics(↑)	Thermal(↑)	Material(↑)	Average(↑)
Before Rewriting Prompt						
CogVideoX (Yang et al., 2024)	5B	0.39	0.55	0.40	0.42	0.45
Kling	-	0.45	0.58	0.50	0.40	0.49
After Rewriting Prompt						
CogVideoX (Yang et al., 2024)	5B	0.39	0.62	0.53	0.52	0.52
Kling	-	0.50	0.64	0.61	0.48	0.56

Table 12: **PCA evaluation results with proposed *PhyGenEval* in videos after VEnhancer**. The results indicate that employing VEnhancer fails to enhance the model's comprehension of physical commonsense.

Model	Size	Mechanics(↑)	Optics(↑)	Thermal(↑)	Material(↑)	Average(↑)
Vchitect 2.0	2B	0.41	0.56	0.44	0.37	0.45
Vchitect 2.0 (Venhancer)	2B	0.41	0.56	0.42	0.38	0.45

does help the models generate images aligned with physical laws, but it is still far from resolving the issues highlighted by *PhyGenBench*. Both CogVideoX 5b and Kling exhibit some growth, but even for Kling, it only achieves a score of 0.56. This demonstrates that current models still severely lack the ability to accurately render physical scenes, and this deficiency cannot be easily resolved through simple prompt rewriting. To illustrate this issue more clearly, as shown in Figure 9, our qualitative analysis shows that rewriting prompts can only address simple issues (e.g., flame color reactions), but remains ineffective for more complex physical processes (e.g., egg breaking, stone sinking).

The robustness of *PhyGenBench* and *PhyGenEval*. VEnhancer (He et al., 2024a) is a generative space-time enhancement framework that improves existing videos by adding spatial details and synthetic motion in the temporal domain. After enhancement by VEnhancer, Vchitect2.0 shows significant improvement on VBench, even surpassing Kling. However, VEnhancer only enhances the visual quality of videos (e.g., making them more coherent and clear) without addressing the model's poor understanding of physical commonsense.

As shown in Table 12, Vchitect enhanced by VEnhancer still scores similarly to the original version on *PhyGenBench*. We calculate a high Spearman coefficient of 0.86 between model scores on *PhyGenBench* before and after VEnhancer enhancement. This indicates that *PhyGenEval* primarily focuses on physical correctness and is robust to other factors affecting visual quality. Furthermore, it demonstrates that even if a model can generate videos with better general quality (e.g., ranking higher on VBench), it doesn't necessarily imply a better understanding of physical common sense. This highlights the distinction between *PhyGenBench* and benchmarks like VBench that evaluate video quality.

F Computaional Resources

Table 13: Resource consumption of models used in PhyGenEval.

Model	Batch Size	Resources	Times	Memory Utilization Per GPU
GPT-4o(stage2)	8	USD 1.4	5min	-
GPT-4o(stage3)	8	USD 3.1	5min	-
LLaVA-Next-Interleave-7B	1	1 x A100-80GB	2min	20408MiB
VQAScore	3	1 x A100-80GB	10min	72726MiB
InternVideo	1	1 x A100-80GB	1min	7766MiB

Although our method involves different stages, it remains straightforward. Table 13 provides a summary of the resource consumption, showing that the entire evaluation process can be completed quickly and at low cost.

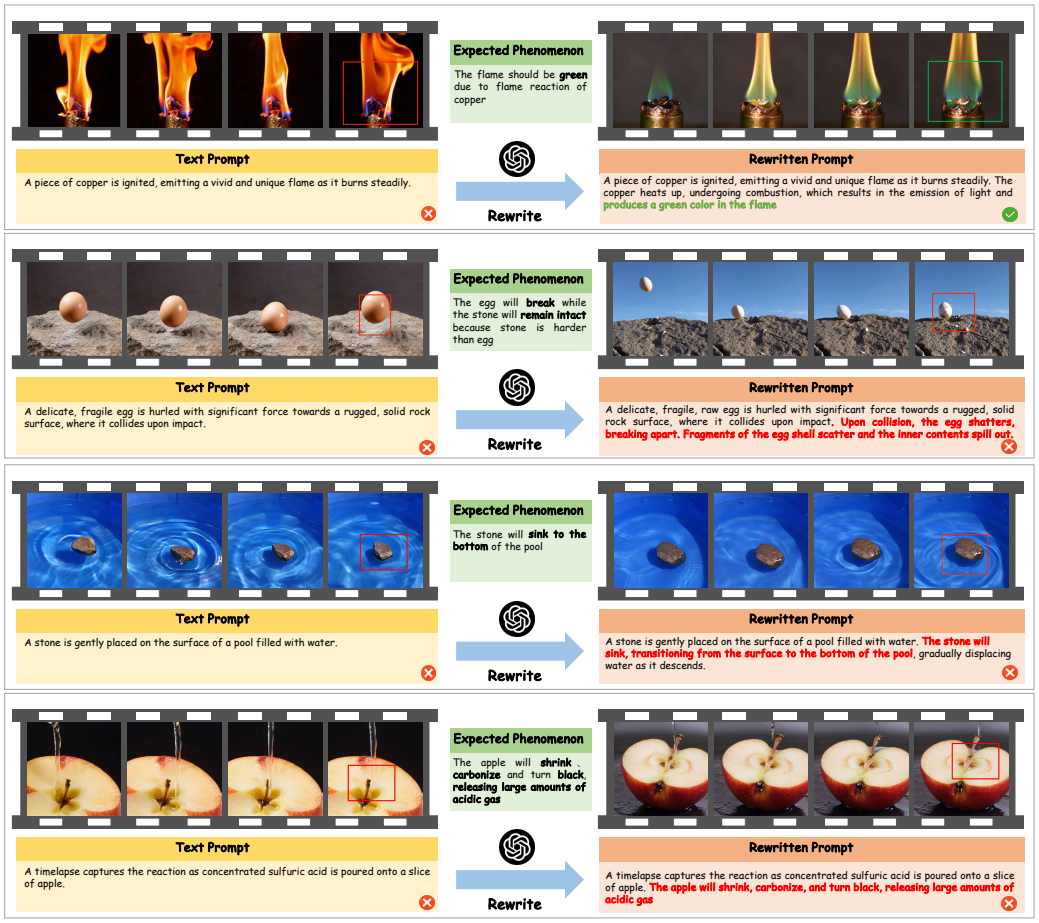


Figure 9: The qualitative comparison of effects before and after using rewritten prompts. The results indicate that rewriting prompts addresses only a few basic issues (such as flame color reactions), while the majority of problems remain unsolved.

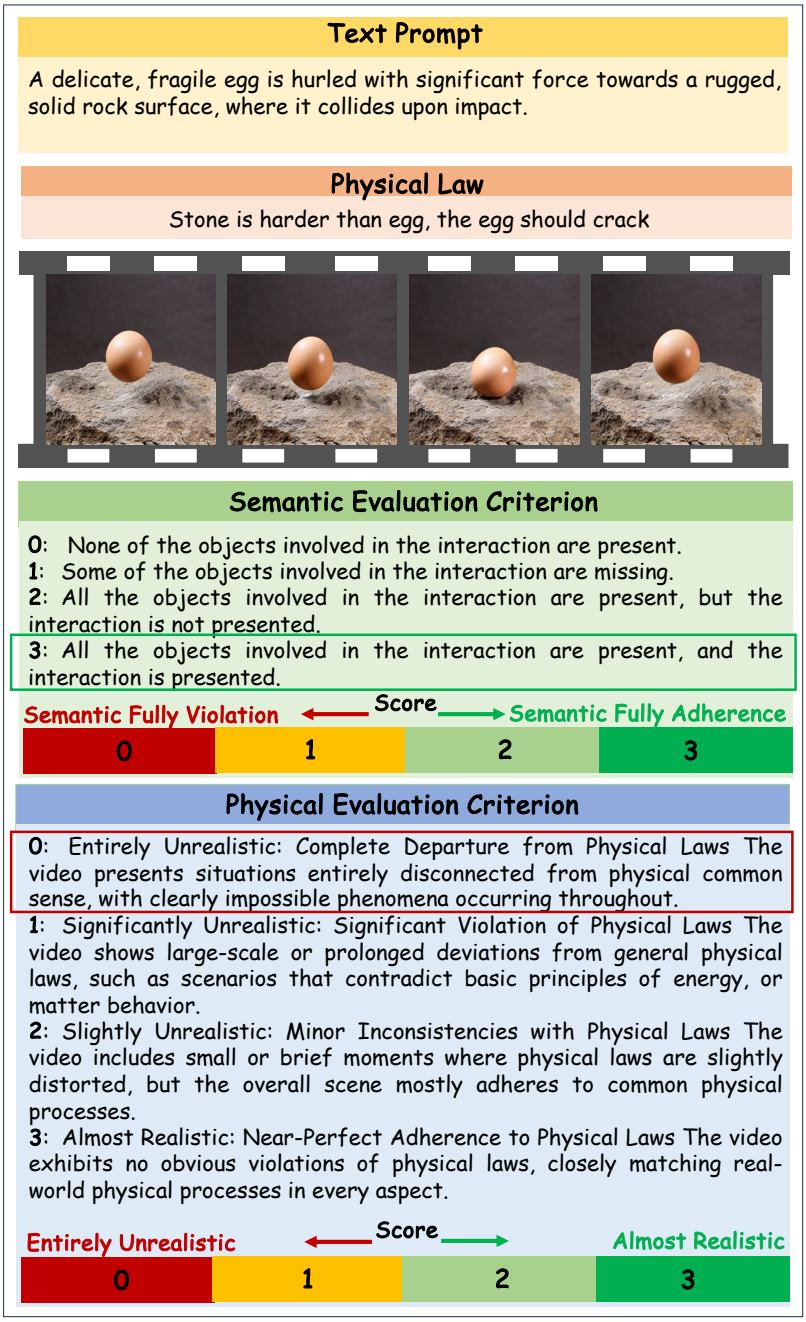


Figure 10: Detailed diagram of the human evaluation process. We ask the annotators to score the semantic alignment and physical commonsense alignment of the video according to the scoring criteria in the figure.