

PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation

Qiyao Xue, Xiangyu Yin, Boyuan Yang and Wei Gao
University of Pittsburgh
 {qix63, eric.yin, by.yang, weigao}@pitt.edu

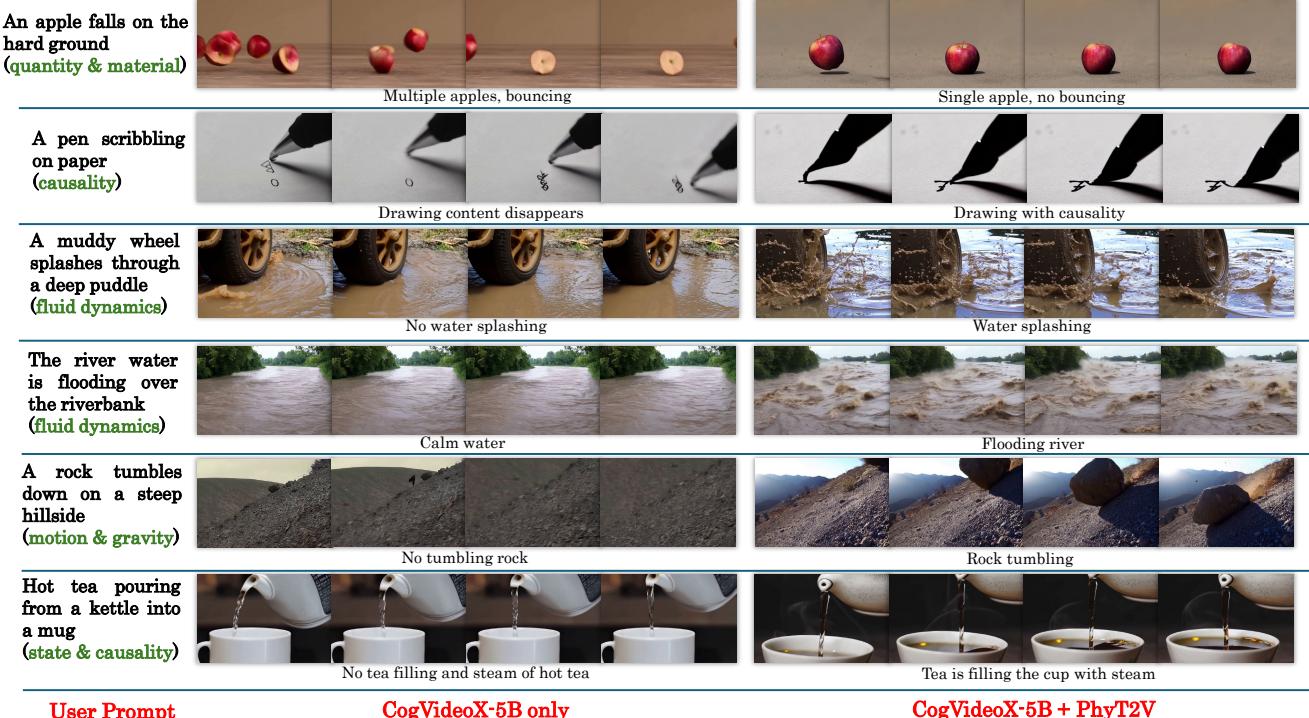


Figure 1. *Left:* videos generated by the current text-to-video generation model (CogVideoX-5B [50]) cannot adhere to the real-world physical rules (described in brackets following the user prompt). *Right:* our method PhyT2V, when applied to the same model, better reflects the real-world physical knowledge.

Abstract

Text-to-video (T2V) generation has been recently enabled by transformer-based diffusion models, but current T2V models lack capabilities in adhering to the real-world common knowledge and physical rules, due to their limited understanding of physical realism and deficiency in temporal modeling. Existing solutions are either data-driven or require extra model inputs, but cannot be generalizable to out-of-distribution domains. In this paper, we present PhyT2V, a new data-independent T2V technique that expands the current T2V model’s capability of video generation to out-of-distribution domains, by enabling chain-of-thought and step-back reasoning in T2V prompting. Our experiments show that PhyT2V improves existing T2V models’ adherence to real-world physical rules by 2.3x, and achieves 35% improvement compared to T2V prompt enhancers.

1. Introduction

Text-to-video (T2V) generation has recently marked a significant breakthrough of generative AI, with the advent of transformer-based diffusion models such as Sora [3], Pika [17] and CogVideoX [51] that can produce videos conditioned on textual prompts. These models demonstrate astonishing capabilities of generating complex and photorealistic scenes, and could even make it difficult for humans to distinguish between real-world and AI-generated videos, in the aspect of individual video frames’ quality [1, 37].

On the other hand, as shown in Figure 1 - Left, current T2V models still have significant drawbacks in adhering to the real-world common knowledge and physical rules, such as quantity, material, fluid dynamics, gravity, motion, collision and causality, and such limitations fundamentally prevent current T2V models from being used for real-world

simulation [7, 19, 31]. Enforcement of real-world knowledge and physical rules in T2V generation, however, is challenging because it requires the models’ understandings of not only individual objects but also how these objects move and interact with each other. Further, unlike generating static images, T2V generation requires frame-to-frame consistency in object appearance, shape, motion, lighting and other dynamics [11]. Current T2V models often lack such temporal modeling, especially over long sequences [20], and the generated videos often contain flickering, inconsistent motion and object deformations across frames [26].

Most of the existing solutions to these challenges are *data-driven*, by using large multimodal T2V datasets that cover different real-world domains to train the diffusion model [10, 12, 41, 49]. However, these solutions heavily rely on the volume, quality and diversity of datasets [42, 51]. Since real-world common knowledge and physical rules are not explicitly embedded in the T2V generation process, the quality of video generation would largely drop in out-of-distribution domains that are not covered by the training dataset, and the generalizability of T2V models is limited due to the vast diversity of real-world scenario domains. Alternatively, researchers also use the existing 3D engines (e.g., Blender [8], Unity3D [36] and Unreal [16]) or mathematical models of edge and depth maps [26–28] to inject real-world physical knowledge into the T2V model, but these approaches are limited to fixed physical categories and patterns such as predefined objects and movements [26, 49], similarly lacking generalizability.

To achieve generalizable enforcement of physics-grounded T2V generation, we propose a fundamentally different approach: instead of expanding the training dataset or further complicating the T2V model architecture, we aim to expand the current T2V model’s capability of video generation from in-distribution to out-of-distribution domains, by embedding real-world knowledge and physical rules into the text prompts with sufficient and appropriate contexts. To avoid ambiguous and unexplainable prompt engineering [9, 32, 33], our basic idea is to enable chain-of-thought (CoT) and step-back reasoning in T2V generation prompting, to ensure that T2V models follow correct physical dynamics and inter-frame consistency by applying step-by-step guidance and iterative refinement.

Based on this idea, this paper presents **Physical-grounded Text-to-Video (PhyT2V)**, a new T2V technique that harnesses the natural language reasoning capabilities of well-trained LLMs (e.g., ChatGPT-4o), to facilitate CoT and step-back reasoning as described above. As shown in Figure 2, such reasoning is iteratively conducted in PhyT2V, and each iteration autonomously refines both the T2V prompt and generated video in three steps. In Step 1, the LLM analyzes the T2V prompt to extract objects to be shown and physical rules to follow in the video via in-context learning. In Step 2, we first use a video captioning model to translate

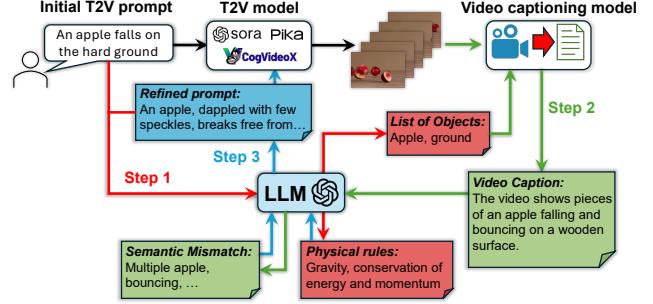


Figure 2. One iteration of video and prompt self-refinement in PhyT2V. Such self-refinement will be iteratively conducted in multiple rounds until the quality of generated video is satisfactory.

the video’s semantic contents into texts according to the list of objects obtained in Step 1, and then use the LLM to evaluate the mismatch between the video caption and current T2V prompt via CoT reasoning. In Step 3, the LLM refines the current T2V prompt, by incorporating the physical rules summarized in Step 1 and resolving the mismatch derived in Step 2, through step-back prompting. The refined T2V prompt is then used by the T2V model again for video generation, starting a new round of refinement. Such iterative refinement stops when the quality of generated video is satisfactory or the improvement of video quality converges.

For physical-grounded video generation performance, we further evaluated PhyT2V by applying it onto multiple SOTA T2V models, by using ChatGPT4 o1-preview [18] for LLM reasoning and Tarsier [39] as the video captioning model. We used two major T2V prompt datasets that cover 7 different real-world domains, and compared PhyT2V with the most competitive baselines of prompt enhancers. Our main findings are as follows:

- PhyT2V is highly effective. Without involving any model retraining efforts on any auxiliary model inputs, PhyT2V can improve the adherence of the existing T2V models’ generated videos to real-world physical rules by up to 2.3x, by only refining the text prompts to the T2V model.
- PhyT2V is high generic. It can result in significant improvement of video quality in a large diversity of real-world domains, covering solid, liquid, mechanics, optics, thermal, etc. It is fully data independent, and its prompting templates can be applied to any existing T2V models with different architectures and input formats.
- Based on LLM-guided reasoning and self-refinement, PhyT2V is fully automated and involve the minimum amount of engineering and manual efforts.

2. Related Work and Motivation

2.1. T2V Generation Models

Early T2V techniques generate video frames from text-to-image model outputs with temporal extensions [35], but cannot maintain temporal consistency and coherence over time, often producing visually appealing but temporally dis-

connected outputs. Diffusion Transformers (DiT) [30] improved such consistency with a transformer backbone capable of capturing more complex temporal dynamics and relationships across frames through attention mechanism and long-range dependency modeling [42, 51]. Based on the DiT architecture, recent T2V models, such as OpenSora [53] and VideoCrafter [4], demonstrated that T2V generation can be further improved by in-context learning when provided with sufficient contextual information [44].



Figure 3. Examples of videos generated from in-distribution and out-of-distribution prompts, using the CogVideoX-5B model

However, as shown in Figure 3, although these T2V models demonstrate strong capabilities in video generation when dealing with prompts aligned with the distributions found in the training data, they encounter significant challenges with out-of-distribution prompts that are not covered by training data¹. In such cases, the outputs often contain physical illusions or artifacts, reflecting the model’s limitations in generating realistic and coherent video contents under unfamiliar conditions. Such limitations can be addressed by enlarging the training datasets, improving T2V model architectures or developing new mechanisms for adaptation and error correction [43, 45], but these approaches are all prompt-specific and lack generalizability.



Figure 4. A video generated by enhancing the out-of-distribution prompt “Whisking egg into milk for scramble” in Figure 3

On the other hand, as shown in Figure 4, recent research has demonstrated that the quality of video generation with an out-of-distribution prompt can be improved by refining the prompt with sufficient and appropriate details [11, 51]. These findings motivate our design of PhyT2V that embeds contexts of real-world knowledge and physical rules into

¹In Figure 3, the in-distribution prompts are picked from the ones listed in [50], and the out-of-distribution prompts are our crafted ones for similar scenarios as the in-distribution prompts.

T2V prompts, to guide the T2V process for better physical accuracy and temporal alignment. The existing works, however, could still fail when tackling more intricate scenarios such as multi-object interactions, because the T2V model lacks an efficient feedback mechanism to learn how the generated video deviates from the real-world knowledge and physical rules. Researchers suggest to provide such feedback with extra input modalities to T2V models such as sampled video frames, depth map or scribble images [44, 52], but incur significant amounts of extra computing overhead and cannot be generalizable. Instead, in our design of PhyT2V, we aim to fully automate the feedback with only text prompts, and enable iterative feedback for the optimum video quality.

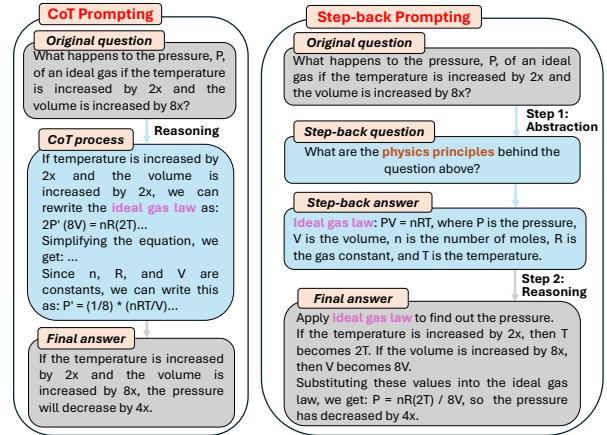


Figure 5. Examples of CoT and step-back reasoning

2.2. Using LLM in T2V Generation

LLMs with strong capabilities in natural language processing (NLP) have been a natural choice for prompt refinement in text-to-image and text-to-video generation, and existing work utilized LLMs to interpret text prompts and orchestrate the initial layout configurations [13, 14, 23–25, 46, 48, 54]. However, since current LLMs lack inherent understandings of the real-world physical laws, using LLMs with simple instructions usually result in videos that appear visually coherent but lack accurate physical realism, particularly when generating scenes with complex object interactions. Furthermore, these approaches frequently rely on static prompts or simple iterative refinements based on bounding box and segmentation map, which may capture basic visual attributes but fail to adapt to nuanced changes that require continuous physical modeling and adjustment.

An effective approach to addressing these limitations and providing effective feedback for prompt refinement is to explicitly trigger in-context learning and reasoning in LLM. For example, as shown in Figure 5, CoT reasoning deconstructs complex prompts into stepwise logical tasks, and hence ensures a precise scheduling path to align generated content with the input prompt. However, CoT reasoning, in some cases, could make errors in some intermediate steps,

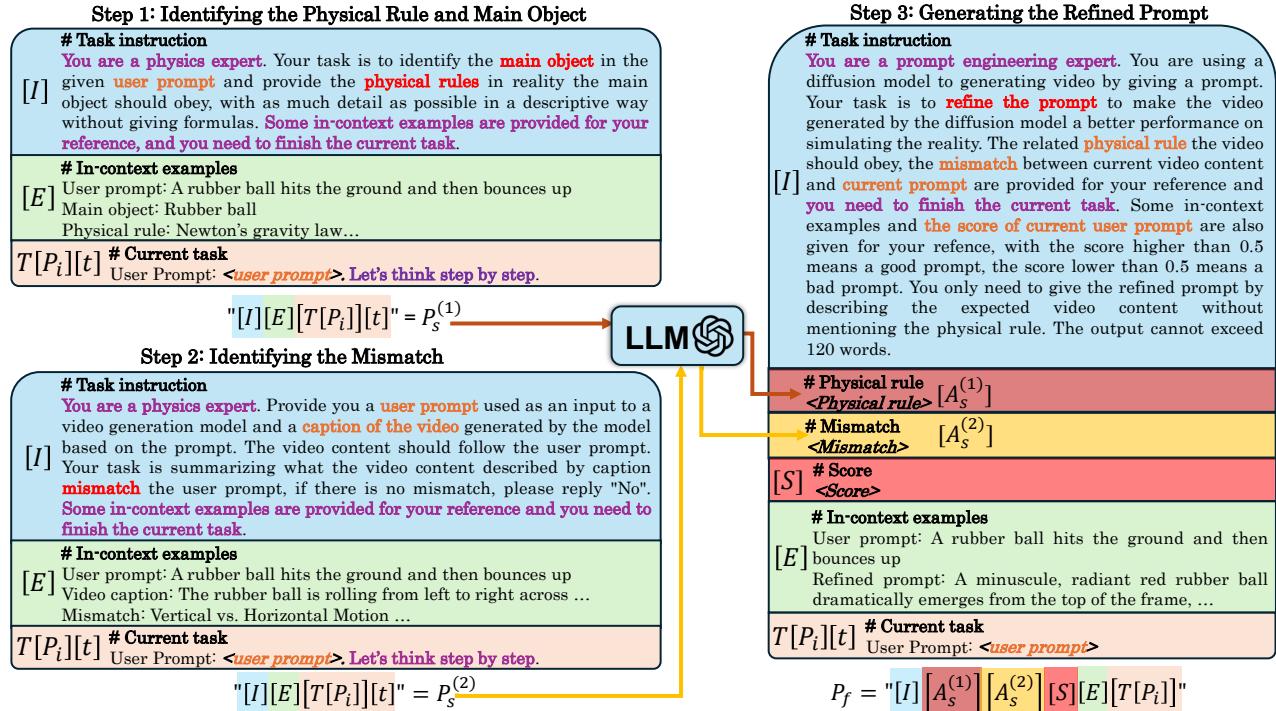


Figure 6. Our design of PhyT2V, illustrated by one round of video refinement consisting of three steps. Texts in brown are inputs from previous step. Texts in red are outputs to the next step; Texts in purple are prompts to trigger LLM reasoning

and step-back prompting can address this limitation by further deriving the step-back question at a higher level of abstraction and hence avoiding confusions and vagueness. In our design of PhyT2V, we will utilize such LLM reasoning to analyze the inconsistency of the generated video to real-world common knowledge and physical rules, and use the reasoning outcome as feedback for T2V prompt refinement.

The Chain-of-Thought (CoT) method is suitable for single data mode processing because it emphasizes linear decomposition and step-by-step reasoning, and it is especially effective for data processing flows that do not require cross-modal synchronization or interaction. However, multimodal data processing involves the fusion of data from different modalities and complex synchronization requirements, which cannot be completed through simple linear decomposition and requires frequent cross-modal information interaction and parallel processing, which exceeds the linear reasoning ability of the CoT method, resulting in limited performance in multimodal tasks. This is the reason why it is hard to directly apply CoT reasoning in T2V process itself as it requires multimodal alignment between the next and video modality. This also motivates us to adopt video captioning and use the video caption in the reasoning process, so that we can conduct CoT and step-back reasoning only in the text domain.

3. Method

In this section, we present details of our PhyT2V design. In principal, PhyT2V's refinement of T2V generation is an

iterative process consisting of multiple rounds. In each round, as shown in Figure 6, the primary objective of our PhyT2V design is to guide a well-trained LLM (e.g., ChatGPT-4o) to generate a refined prompt that enables the pre-trained T2V model to generate videos that better match the given user prompt and real-world physical rules, and the refined prompt will be iteratively used as the new user prompt in the next round of refinement.

Each round of refinement is structured around decomposing the complex refinement problem into a series of simpler subproblems, more specifically, two parallel subproblems and one final subproblem. The two parallel subproblems are: *Step 1*) identifying the relevant physical rules that the generated video should follow based on the user prompt, and *Step 2*) identifying semantic mismatches between the user prompt and the generated video. Based on the knowledge about physical rules and semantic mismatches, the final subproblem (*Step 3*) generates the refined prompt to better adhere to the physical rules and resolve the mismatches.

To ensure proper identification in the parallel subproblems and prompt generation in the final subproblem, the core of PhyT2V design is two types of LLM reasoning processes within the prompt enhancement loop: the *local CoT reasoning* for individual subproblems and *global step-back reasoning* for the overall prompt refinement problem.

Local CoT reasoning is executed within the prompt for each subproblem, to prompt the LLM to generate a detailed reasoning chain in its latent embedding space [38]. Addressing the parallel subproblems facilitates LLM with a

more concentrated attention on prerequisites of prompt refinement, enabling a deeper comprehension of the physical laws that govern the video content as well as the identification of discrepancies between the generated video and the user prompt. The outcomes derived from these parallel subproblems reflect the language model’s abstraction in step-back reasoning on the overarching prompt refinement.

Global step-back reasoning: To integrate various subproblems into a coherent framework for prompt and video refinement, one intuitive approach involves employing CoT reasoning across these subproblems, allowing the LLM to engage in self-questioning. However, this method may lead to the risk of traversing incorrect reasoning pathways. Instead, we apply global step-back reasoning across subproblems, by using a self-augmented prompt to incorporate the LLM-generated responses to high-level questions about physical rules and semantic mismatches in earlier parallel problems, when generating the refined prompt in the final subproblem. In this way, we can improve the correctness of intermediate reasoning steps in CoT reasoning, and enable consistent improvement across steps in reasoning.

Both reasoning processes are facilitated through appropriate task instruction prompting tailored to different subproblems. In general, our prompting procedure follows the prompt modeling in [34], which divides task instructions into several components. More details about these components in our design of PhyT2V are elaborated as follows.

Compared to the previous prompt enhancing methods, PhyT2V’s key contribution is to analyze the semantic mismatch between currently generated video and the prompt, as well as refinements based on such mismatch. Previous methods can be formulated as $p' = f_{enhance}(p, \theta)$, where p and p' are the original and enhanced prompts, $f_{enhance}$ is the prompt enhancer, and θ represents parameters or rules guiding the enhancement. In contrast, PhyT2V further involves the additional information about the T2V process, i.e., $p' = f_{enhance}(p, f_{mismatch}(C(V(p)), p), f_{phy}(p), \theta)$, where $f_{phy}(p)$ analyzes the physical rules to be followed given p , $V(p)$ is the currently generated video given prompt p , C is the video captioning model and $f_{mismatch}$ finds the semantic mismatch between $C(V(p))$ and p . The key advantages are: (1) Semantic awareness: the refinement process explicitly incorporates the semantic mismatch to enable targeted T2V improvements; (2) Physical-world knowledge integration: physical rules derived from p enable guided enhancement; (3) Guided reasoning: unlike prior methods that rely solely on templates or embeddings, PhyT2V dynamically adapts prompt refinement to the semantic mismatch.

3.1. Prompting in Parallel Subproblems for Local CoT Reasoning

In both Step 1 and Step 2, the first part of prompt is a task instruction prompt $[I]$ to instruct the LLM to understand the task in the subproblem. $[I]$ is designed with multiple components, each of which corresponds to different functions.

In the first sentence, it provides general guidance to relate the current subproblem to the entire refinement problem, to better condition the subproblem answer. Afterwards, it will include detailed descriptions of the task: identifying the physical rule and main object in Step 1, and identifying the semantic mismatch between the user prompt and caption of the generated video (generated by the video captioning model) in Step 2. It will also contain the requirements about the expected information in LLM’s output. For example, in Step 1, the LLM’s output about the physical rule should be in a descriptive way without giving formulas.

Besides, to ensure proper CoT reasoning, we follow the existing work [22, 40] and provide in-context examples $[E]$ about tasks. To facilitate LLM’s in-context learning [5, 6], $[E]$ is given in the format of QA pairs. That is, instead of fine-tuning a separate LLM checkpoint for each new task, we prompt the LLM with a few input-output exemplars, to demonstrate the task and condition the task’s input-output format to the LLM, to guide the LLM’s reasoning process.

Then, the final part of the prompt, denoted as $[T]$, is the information of the current instance of the task, usually with the current user prompt (P_i) being embedded. As a common practice of CoT reasoning, it also contains the hand-crafted trigger phrase (t), “Let’s think step by step”, to activate the local CoT reasoning in LLM.

3.2. Prompting in the Final Subproblem for Global Step-Back Reasoning

In the final subproblem, we enforce global step-back reasoning, by using the outputs of the two parallel subproblems above, i.e., knowledge about the physical rules and the prompt-video mismatch, as the high-level concepts and facts. Grounded on such high-level abstractions, we can make sure to improve the LLM’s ability in following the correct reasoning path of generating the refined prompt.

Being similar to the prompts used in the two parallel subproblems above, the prompt structure in the final subproblem also contains $[I]$, $[E]$ and $[T]$. Furthermore, to ensure the correct reasoning path, we also provide quantitative feedback to the LLM about the effectiveness of previous round’s prompt refinement. Such effectiveness could be measured by the existing T2V evaluators, which judge the semantic alignment and quality of physical common sense of the currently generated video². For example, the VideoCon-Physics evaluator [2] gives a score ($[S]$) between 0 and 1. If $[S]$ is <0.5 , it indicates that the refined prompt produced in the previous round is ineffective, hence guiding the LLM to take another alternative reasoning path.

Since the prompt in the final subproblem is rich with reasoning and inherently very long-tailed, we removed the trigger prompt $[t]$, to prevent incorporating the information in the final answer unrelated to the user’s initial input prompt.

²This video is generated using the prompt refined in the previous round, and is also used to generate the video caption as the input in Step 2.

3.3. The Stopping Condition

The process of iterative refinement normally continues until the quality of the generated video is satisfactory, measured by the T2V evaluator as described above. Furthermore, the current T2V models naturally have limitations in generating some complicated or subtle scenes. In these cases, it would be difficult, even for PhyT2V, to reach physical realism after multiple rounds of iterations, and PhyT2V’s refinement would stop when the iterations converge, i.e., the improvement of video quality becomes little over rounds.

4. Experiments

Models & Datasets: We applied PhyT2V on several DiT-based open-source T2V models, as listed below, and evaluated how PhyT2V improves the generated videos’ adherence to real-world knowledge and physical rules. We use ChatGPT4 o1-preview [18] as the LLM for reasoning, and Tarsier [39] as the video captioning model. All generated videos last 6 seconds with 10 FPS and resolution of 720×480 . Details of evaluation setup are in Appendix A.

- **CogVideoX [51]:** It generates 10-second videos from text prompts, with 16 FPS and 768×1360 resolution. It offers two model variants with 2B and 5B parameters.
- **OpenSora 1.2 [53]:** As an alternative to OpenAI’s Sora [3], it contains 1.1B parameters and produces videos with 16 seconds, 720p resolution and different aspect ratios.
- **VideoCrafter [4]:** With 1.8B parameters, it can generate both images and videos from text prompts, with 576×1024 resolution and a focus on video dynamics.

Since we target enhancing the T2V models’ capability of generating physics-grounded video contents, we use the following two prompt benchmarks that emphasize physical laws and adherence as the text prompts for T2V:

- **VideoPhy [2]** is designed to assess whether the generated videos follow physical common sense for real-world activities. It consists 688 human-verified captions that describe interactions between various types of real-world objects, including solid and fluid.
- **PhyGenBench [29]**, similarly, allows evaluating the correctness of following physical common sense in T2V generation. It comprises 160 carefully crafted prompts spanning four physical domains, namely mechanics, optics, thermal and material properties. Since the domain of material properties has been covered by VideoPhy, we use the first three domains listed above.

Evaluation metric: We use VideoCon-Physics evaluator provided with VideoPhy [2], to measure how the generated video adheres to physical common sense (PC) and achieves semantic adherence (SA). The PC metric evaluates whether the depicted actions and object’s state follow the real-world physics laws. The SA metric measures if the actions, events, entities and their interactions specified in the prompt are present. Both metrics yield binary outputs: 1 indicates adherence and 0 indicates otherwise. On each T2V model and

	Round	1	2	3	4
CogVideoX-5B [51]	PC	0.26	0.32	0.39	0.42
	SA	0.48	0.52	0.56	0.59
CogVideoX-2B [51]	PC	0.13	0.19	0.27	0.29
	SA	0.22	0.12	0.40	0.42
OpenSora [53]	PC	0.17	0.29	0.27	0.31
	SA	0.29	0.38	0.44	0.47
VideoCrafter [4]	PC	0.15	0.25	0.29	0.33
	SA	0.24	0.38	0.44	0.49

Table 1. The quality of videos generated by different T2V models using the VideoPhy prompt dataset, over multiple rounds of iterative refinement in PhyT2V

	Round	1	2	3	4
CogVideoX-5B [51]	PC	0.28	0.32	0.38	0.42
	SA	0.22	0.35	0.36	0.38
CogVideoX-2B [51]	PC	0.16	0.19	0.24	0.27
	SA	0.15	0.29	0.33	0.35
OpenSora [53]	PC	0.21	0.25	0.24	0.26
	SA	0.23	0.28	0.29	0.30
VideoCrafter [4]	PC	0.20	0.24	0.32	0.36
	SA	0.27	0.33	0.37	0.42

Table 2. The quality of videos generated by different T2V models using the PhyGenBench prompt dataset, over multiple rounds of iterative refinement in PhyT2V

dataset, the binary outputs from all prompts are averaged.

In addition, we also evaluated PhyT2V using the widely used VBench metrics and benchmarks [15], which allow comprehensive evaluations of the generated video in multiple aspects, including video quality, video-condition consistency, prompt following and human preference alignment.

Baselines: For fair comparison, we only use the existing T2V prompt enhancers as baselines, and other existing work with extra inputs to T2V models [7, 19, 26, 27, 31] are not applicable. We involve two prompt enhancers: 1) Directly using the existing LLM, particularly ChatGPT4, as the prompt enhancer [28, 47]; 2) Promptist [21], which uses reinforcement learning to automatically refine and enhance prompts in the model-preferred way.

4.1. Improvement of the Generated Video Quality

As shown in Table 1 and 2, when PhyT2V is applied to different T2V models, it can significantly improve the generated video’s adherence to both the text prompt itself and the real-world physical rules, compared to the videos generated by vanilla T2V models (i.e., in Round 1 of PhyT2V’s refinement). In particular, such improvement is the most significant on the CogVideoX-2B model, where PC improvement can be up to 2.2x and SA improvement can be up to 2.3x. On all the other models, PhyT2V can also reach noticeable improvement, ranging from 1.3x to 1.9x.

Meanwhile, results in Table 1 and 2 showed that

	CogVideoX-5B	OpenSora
ChatGPT 4 [28]	PC	0.33
	SA	0.41
Promptist [21]	PC	0.25
	SA	0.39

Table 3. The quality of videos generated by enhancing the prompts in the VideoPhy dataset using different prompt enhancers

	CogVideoX-5B	OpenSora
ChatGPT 4 [28]	PC	0.27
	SA	0.23
Promptist [21]	PC	0.32
	SA	0.24

Table 4. The quality of videos generated by enhancing the prompts in the PhyGenBench dataset using different prompt enhancers

PhyT2V’s process of iterative refinement converge quickly and only takes few rounds. Most improvement of video quality happens in the first two rounds, and little improvement can be observed in the fourth round. Hence, in practice, we believe that 3-4 iterative rounds would be sufficient.

Furthermore, as shown in Table 3 and 4, PhyT2V also largely outperforms the existing prompt enhancers by at least 35%, when being applied to CogVideoX-5B and OpenSora models. In particular, ChatGPT 4, when being used as the prompt enhancer, delivers better performance than Promptist due to its stronger language processing capabilities, but still cannot ensure physics-grounded T2V, due to the lack of explicit reasoning on text-to-video alignment.

Our evaluation results on the VBench metrics are shown in Figure 7, where numbers in Round 1 are the T2V model’s original performance in current VBench leaderboard, and iterative prompt refinements by PhyT2V in Round 2 & 3 noticeably improve the performance in many dimensions. In particular, large improvements are noted in most dimensions of Video-Condition Consistency, showing that PhyT2V improves T2V model’s adherence to prompts and real-world physical rules underlying the prompts.

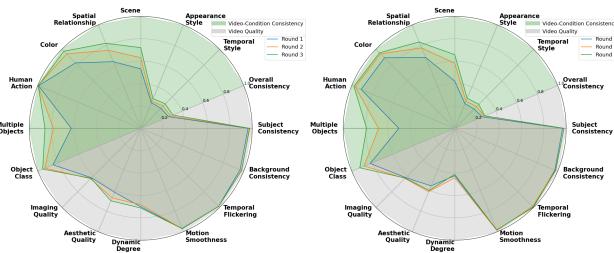


Figure 7. PhyT2V VBench evaluation results with CogVideoX-5B (left) and OpenSora (right)

4.2. Different Domains of Physical Rules

We also conducted in-depth analysis on PhyT2V’s performance on improving the generated video’s quality in different domains of real-world physical rules, using the

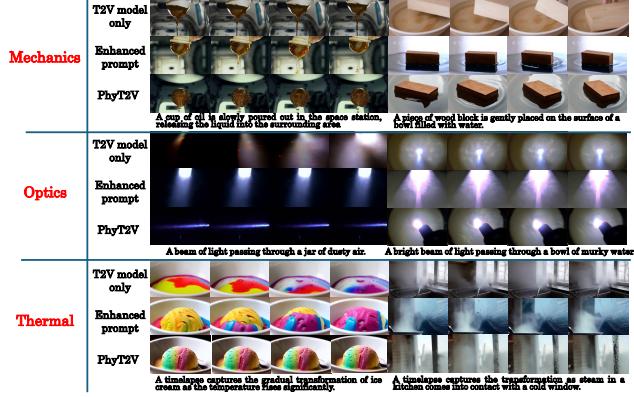


Figure 8. Examples of videos generated using different categories of prompts in the PhyGenBench dataset

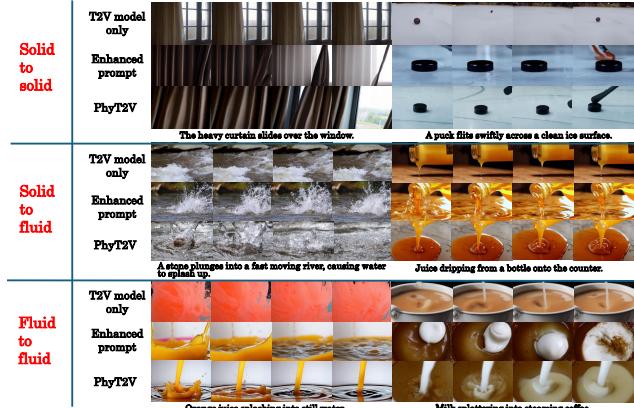


Figure 9. Examples of videos generated using different categories of prompts in the VideoPhy dataset

CogVideoX-5B as the T2V model and ChatGPT 4 as the prompt enhancer. As shown in Table 5 and 6, PhyT2V achieves large improvements in most domains of physical rules. Especially in domains where physical interaction between objects are more subtle and difficult to be precisely captured, such as interaction with fluids and thermal-related scene changes, such improvements will be even higher.

These improvements are also exemplified with sample videos and their related input prompts in Figure 9 and Figure 8. With LLM reasoning and iterative refinement, PhyT2V can largely enhance the T2V model’s capability when encountering out-of-distribution prompts, by providing correct and sufficient contexts to ensure that the T2V model’s video generation correctly capture the key objects and interaction between objects. For example, when the prompt of “juice dropping from a bottle onto the counter”, PhyT2V correctly depicts the juice’s slow diffusion on the counter. More examples can be found in Appendix B.

4.3. Ablation Study

We conduct an ablation study to evaluate the necessity of both the physical rule reasoning (Step 1) and the mismatch reasoning (Step 2) within our PhyT2V workflow, by removing one of these steps from the refinement process to assess

	CogVideoX-5B				CogVideoX-2B				OpenSora				VideoCrafter				
Round	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
Solid-Solid	PC	0.21	0.28	0.34	0.32	0.09	0.13	0.14	0.22	0.12	0.27	0.29	0.30	0.19	0.22	0.27	0.28
	SA	0.24	0.48	0.49	0.47	0.18	0.25	0.36	0.33	0.16	0.34	0.37	0.35	0.24	0.40	0.45	0.47
Solid-Fluid	PC	0.22	0.27	0.28	0.30	0.11	0.18	0.28	0.27	0.17	0.21	0.24	0.25	0.18	0.24	0.25	0.26
	SA	0.39	0.54	0.60	0.61	0.29	0.43	0.44	0.43	0.16	0.40	0.41	0.36	0.34	0.43	0.48	0.52
Fluid-Fluid	PC	0.57	0.59	0.63	0.62	0.34	0.38	0.35	0.36	0.15	0.32	0.29	0.31	0.33	0.41	0.53	0.51
	SA	0.41	0.57	0.59	0.67	0.27	0.42	0.39	0.44	0.31	0.44	0.45	0.46	0.32	0.42	0.49	0.51

Table 5. The improvement of generated video quality in different categories of physical rules in the VideoPhy prompt dataset

	CogVideoX-5B				CogVideoX-2B				OpenSora				VideoCrafter				
Round	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
Mechanics	PC	0.19	0.25	0.34	0.35	0.12	0.16	0.18	0.24	0.11	0.13	0.17	0.22	0.14	0.23	0.29	0.28
	SA	0.21	0.28	0.29	0.32	0.11	0.18	0.19	0.22	0.19	0.21	0.27	0.32	0.20	0.24	0.28	0.35
Optics	PC	0.22	0.35	0.41	0.39	0.22	0.25	0.29	0.28	0.24	0.26	0.25	0.25	0.22	0.21	0.27	0.32
	SA	0.27	0.42	0.39	0.44	0.23	0.34	0.37	0.39	0.26	0.31	0.29	0.30	0.22	0.28	0.35	0.39
Thermal	PC	0.33	0.35	0.35	0.35	0.13	0.15	0.15	0.14	0.27	0.30	0.31	0.33	0.25	0.28	0.26	0.28
	SA	0.22	0.36	0.43	0.45	0.12	0.16	0.24	0.27	0.23	0.25	0.37	0.36	0.25	0.37	0.41	0.43

Table 6. The improvement of generated video quality in different categories of physical rules in the PhyGenBench prompt dataset

its impact on physical-grounded video generation.

Physical rule reasoning (Step 1). As shown in Figure 10, the Step 1 of physical rule reasoning significantly enhances the T2V process by providing a more detailed and coherent description of the principal object’s physical status, such as motion, states and deformation (red texts in Figure 10), all grounded in relevant physical laws. By anchoring the prompt in established physical rules, this step also help avoid unnecessary texts (brown texts in Figure 10) and vague physical rule descriptions (purple texts in Figure 10), hence achieving a higher PC score.

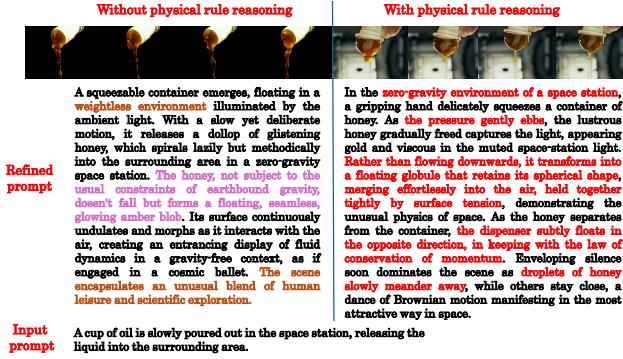


Figure 10. Ablation study on Step 1 of physical rule reasoning

Mismatch reasoning (Step 2). The Step 2 of mismatch reasoning addresses details that may have been overlooked in the previous iteration of the generated video as shown in Figure 11. This step plays a critical role in the iterative refinement process by identifying and correcting discrepancies between expected and observed outputs. By enhancing the model’s focus on the principal object, the mismatch reasoning step reduces the likelihood of losing attention to

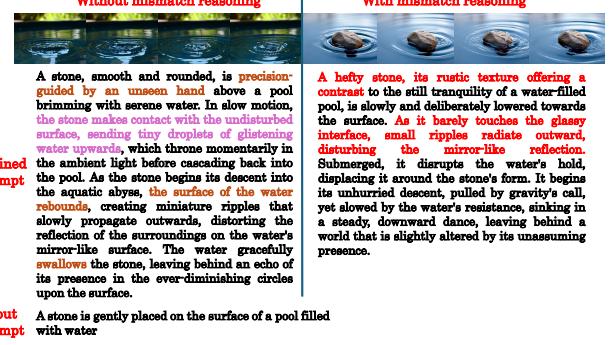


Figure 11. Ablation study on Step 2 of mismatch reasoning

important features (brown and purple texts in Figure 11), improving the fidelity and relevance of generated video content (red texts in Figure 11) towards a higher SA score.

Overall, our study shows that both reasoning steps are integral to PhyT2V, contributing to a more robust and semantically-aligned generation of refined prompts in Step 3. More detailed ablation studies are in Appendix C.

5. Conclusion

In this paper, we present PhyT2V, a novel data-independent T2V generation framework designed to enhance the generalization capability of existing T2V models to out-of-distribution domains. By incorporating CoT reasoning and step-back prompting, PhyT2V systematically refines T2V prompts to ensure adherence to real-world physical principles without necessitating additional model retraining or reliance on additional conditions. Evaluation results indicate that PhyT2V achieves a 2.3x enhancement in physical realism compared to baseline T2V models and outperforms state-of-the-art T2V prompt enhancers by 35%.

Acknowledgments

We thank the reviewers and the area chair for their insightful comments and feedback. This work was supported in part by National Science Foundation (NSF) under grant number IIS-2205360, CCF-2217003, CCF-2215042, and National Institutes of Health (NIH) under grant number R01HL170368.

References

- [1] Luke Auburn. Ai video generation expert discusses the technology’s rapid advances—and its current limitations, 2024. [1](#)
- [2] 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. [5](#) [6](#)
- [3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. [1](#) [6](#)
- [4] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. [3](#) [6](#) [11](#)
- [5] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. [5](#)
- [6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, 2024. [5](#)
- [7] Theodoros Dounas and Alexandros Sigalas. Blender, an open source design tool: Advances and integration in the architectural production pipeline. *Aristoteleio University of Thessaloniki*, 21:737–744, 2009. [2](#) [6](#)
- [8] Blender Foundation. Upbge: an open-source, 3d game engine forked from the old blender game engine, 2024. [2](#)
- [9] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models, 2023. [2](#)
- [10] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pages 393–411. Springer, 2025. [2](#)
- [11] 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. [2](#) [3](#)
- [12] Kai Huang, Haoming Wang, and Wei Gao. Freezeasguard: Mitigating illegal adaptation of diffusion models via selective tensor freezing. *arXiv preprint arXiv:2405.17472*, 2024. [2](#)
- [13] Kai Huang, Hanyun Yin, Heng Huang, and Wei Gao. Towards green ai in fine-tuning large language models via adaptive backpropagation. *ICLR*, 2024. [3](#)
- [14] Kai Huang, Xiangyu Yin, Heng Huang, and Wei Gao. Modality plug-and-play: Runtime modality adaptation in LLM-driven autonomous mobile systems. In *ACM MobiCom*, 2025. [3](#)
- [15] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. [6](#)
- [16] Epic Games Inc. Unreal engine: The most powerful real-time 3d creation tool, 2024. [2](#)
- [17] Mellis Inc. Pika labs, 2023. [1](#)
- [18] OpenAI Inc. Introducing openai o1-preview, 2024. [2](#) [6](#)
- [19] Marcel Krüger, David Gilbert, Torsten W. Kuhlen, and Tim Gerrits. Game engines for immersive visualization: Using unreal engine beyond entertainment. *PRESENCE: Virtual and Augmented Reality*, 33:31–55, 2024. [2](#) [6](#)
- [20] Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects, 2024. [2](#)
- [21] WeiJie Li, Jin Wang, and Xuejie Zhang. Promptist: Automated prompt optimization for text-to-image synthesis. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 295–306. Springer, 2024. [6](#) [7](#) [11](#)
- [22] Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. *Advances in Neural Information Processing Systems*, 36, 2024. [5](#)
- [23] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. [3](#)
- [24] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023.
- [25] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. [3](#)
- [26] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2025. [2](#) [6](#)
- [27] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1430–1440, 2024. [6](#)

- [28] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning, 2024. 2, 6, 7
- [29] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quan-feng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 6
- [30] 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. 3
- [31] Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*, pages 909–916. Springer, 2016. 2, 6
- [32] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2024. 2
- [33] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Gonçarenc, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. The prompt report: A systematic survey of prompting techniques, 2024. 2
- [34] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 2024. 5
- [35] 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. 2
- [36] Unity Technologies. Unity real-time development platform, 2024. 2
- [37] Stuart A. Thompson. A.i. can now create lifelike videos. can you tell what's real?, 2024. 1
- [38] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022. 4
- [39] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 2, 6, 11
- [40] Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yun-sen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*, 2023. 5
- [41] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens, 2024. 2
- [42] 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. 2, 3
- [43] 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. 3
- [44] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36:8542–8562, 2023. 3
- [45] Zhao Wang, Aoxue Li, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024. 3
- [46] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336, 2024. 3
- [47] Deshun Yang, Luhui Hu, Yu Tian, Zihao Li, Chris Kelly, Bang Yang, Cindy Yang, and Yuxian Zou. Worldgpt: a sora-inspired video ai agent as rich world models from text and image inputs. *arXiv preprint arXiv:2403.07944*, 2024. 6, 11
- [48] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [49] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 2
- [50] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihua Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024. 1, 3, 11
- [51] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen 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. 1, 2, 3, 6
- [52] Wentao Zhang, Junliang Guo, Tianyu He, Li Zhao, Linli Xu, and Jiang Bian. Video in-context learning. *arXiv preprint arXiv:2407.07356*, 2024. 3

- [53] 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, 2024. 3, 6, 11
- [54] Hanxin Zhu, Tianyu He, Anni Tang, Junliang Guo, Zhibo Chen, and Jiang Bian. Compositional 3d-aware video generation with llm director. *arXiv preprint arXiv:2409.00558*, 2024. 3

A. Details of Evaluation Setup

Since our proposed technique of PhyT2V does not involve any efforts of retraining the T2V model, in this section we describe details about our evaluation setup of the LLM inference for CoT and step-back reasoning.

In our evaluations, we use 4 T2V generation models, including CogVideoX-5B [50], CogVideoX-2B [50], Open-Sora [53] and VideoCrafter [4]. We choose to use these models as they are all built with transformer based diffusion model, which enhanced the semantic adherence by using the cross attention mechanism, and hold a high score on both the VideoPhy dataset and PhyGenBench dataset leader board. We use Tarsier [39] as the video captioning model for it achieves state of art on multiple video question answering datasets, which ensure the precise and detail of the video captioning step in our approach.

Since PhyT2V improves the quality of generated videos through text prompt refinement, we use the Promptist [21] and GPT-4o [47] as the prompt enhancers, with the same model hyper-parameter setting for the baselines to maintain the consistency between the baseline and our approach. The diffusion model generated video length is setted as 6 second and 8 frames per second with frame resolution 720×480 .

To fit the maximum token input length of the T2V model, we limit the word length of the refined prompts to 120, by instructing the ChatGPT4 o1-preview model that is used as the LLM for reasoning. To formatting the output and storage in our approach, the ChatGPT4 o1-preview model are instructed to output in JSON format and output results in each step are saved in a CSV file by row. The ChatGPT4 o1-preview is called by using the API with the input prompt separated into system prompt and user prompt to identify the instruction part and the rest of the original prompt for ChatGPT. The first and second step output is embedded to the third step prompt template by replacing the pre-defined place holder.

B. More Examples of Physics-Grounded Videos Generated by PhyT2V

B.1. Examples of Generated Videos in Different Categories of Physical Rules

In this subsection, we present the additional comparison examples of generated videos of CogVideo-5b on the VideoPhy and PhyGenBench dataset, Fig 12, 13, 14 show the additional comparison of video generation result on fluid to fluid, solid to fluid, solid to solid cases on VideoPhy dataset, Fig 15, 16, 17 show the additional comparison of video generation result on force, optics and thermal cases on PhyGenBench dataset.

B.2. Refinement process details

In this subsection, we present the prompt refinement detail by the CogVideo-5b with PhyT2V on the VideoPhy and

PhyGenBench dataset, Fig 20, 19, 18 show the prompt refinement details of fluid to fluid, solid to fluid, solid to solid cases on VideoPhy dataset, Fig 21, 22, 23 show prompt refinement details of force, optics and thermal cases on PhyGenBench dataset.

C. Ablation study details

C.1. Model size

We found that the PhyT2V approach can unleashing more power of physical-grounded video generation on a larger model as the result shown by comparing the CogVideo-2b and CogVideo-5b in Figure 24 .

C.2. Prompt template component

In this section some part of the prompt template component is removed to show the necessity of the corresponding components as shown in Figure 25, 26, 27. Without the role indicator sentence, the generated output content is lake of precise information, without the in-context examples, the GPT can not generated the output in an expected format.

D. Failure cases

PhyT2V may be ineffective in two categories of cases.

First, many T2V models exhibit temporal flickering or inconsistent object trajectories, due to absence of long-term temporal coherence mechanisms in model design. Even if prompts are refined to emphasize smooth temporal transitions or continuous motions, these requirements may not be achieved due to model's limitations.

Second, T2V models are typically trained on large datasets, which often lack samples of rare or complex physical phenomena. Hence, these models struggle in scenarios that are underrepresented in the training data. Even with highly specific prompts, the T2V models may still fail to extrapolate effectively to these underrepresented cases.

For some specific generation content categories, we found that even with the PhyT2V refined several rounds, the diffusion model still failed to precisely generating human body, especially on human hands as shown in Figure 28.

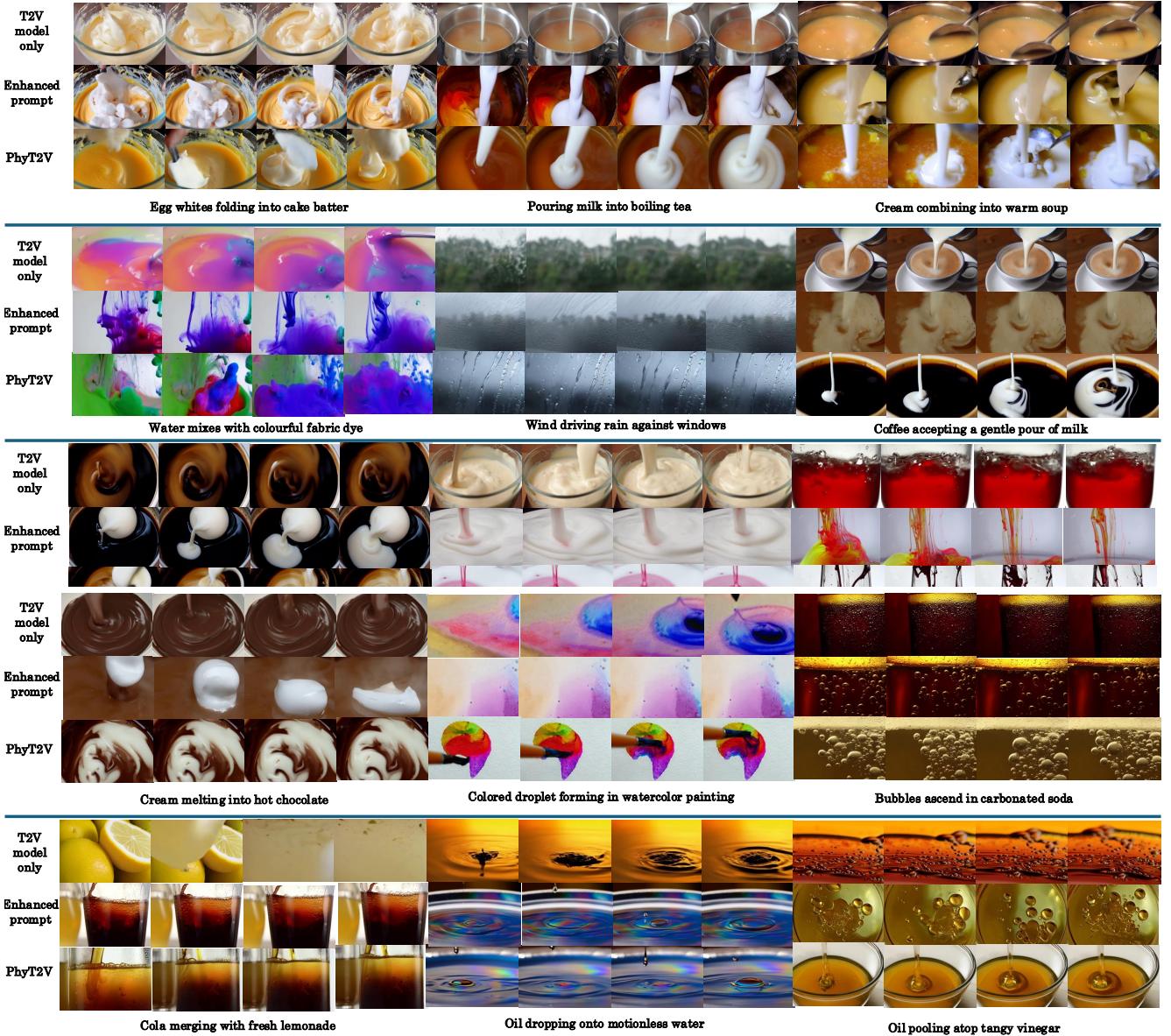
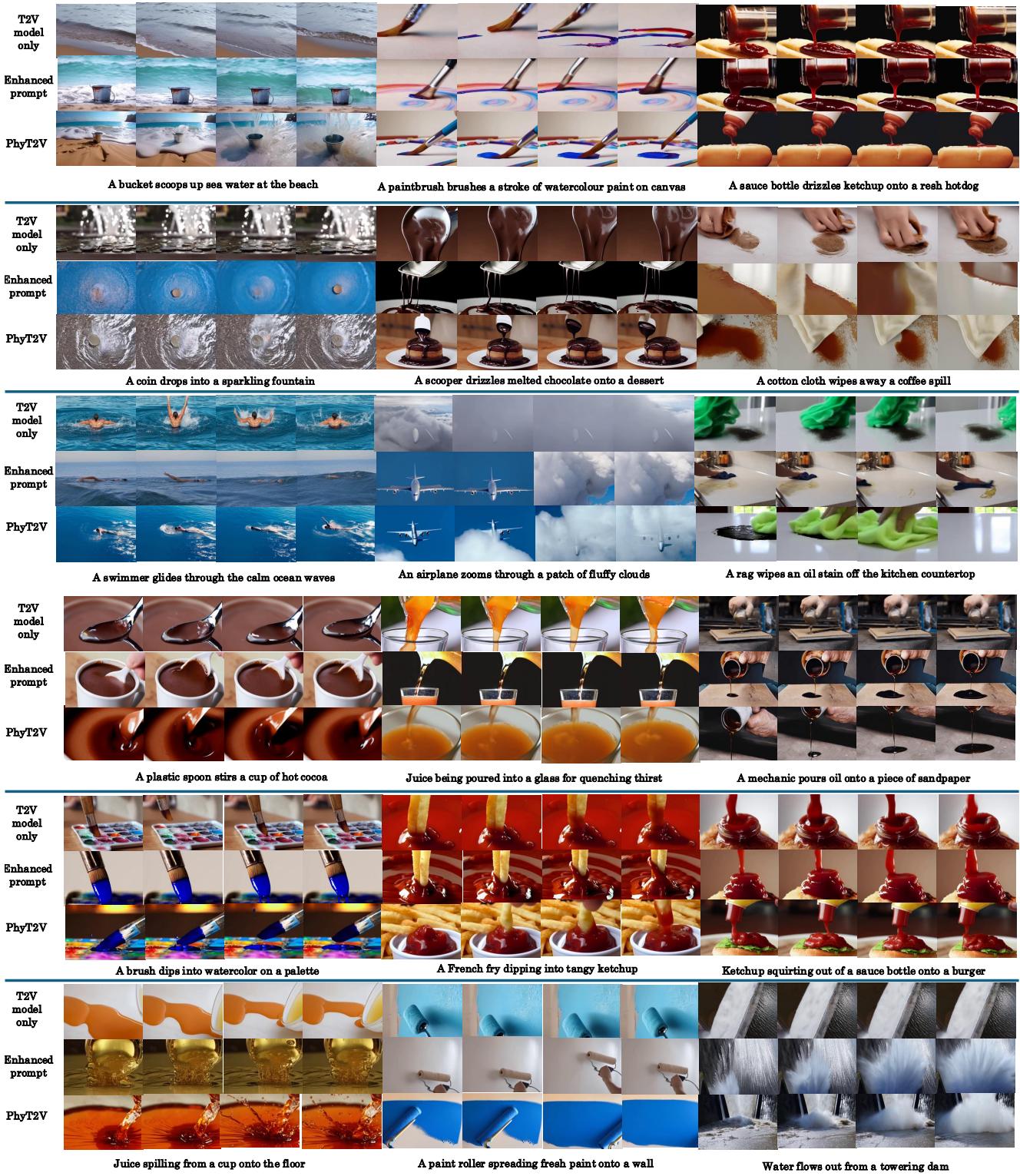


Figure 12. Video generation example on fluid to fluid specific prompt in VideoPhy dataset



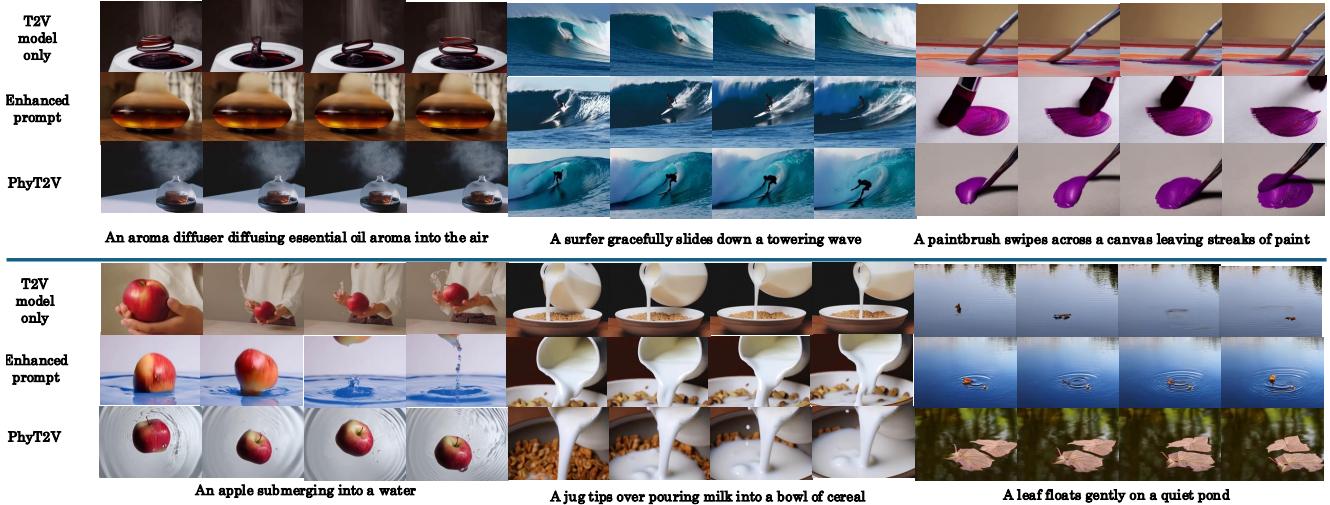


Figure 13. Video generation example on solid to fluid specific prompt in VideoPhy dataset

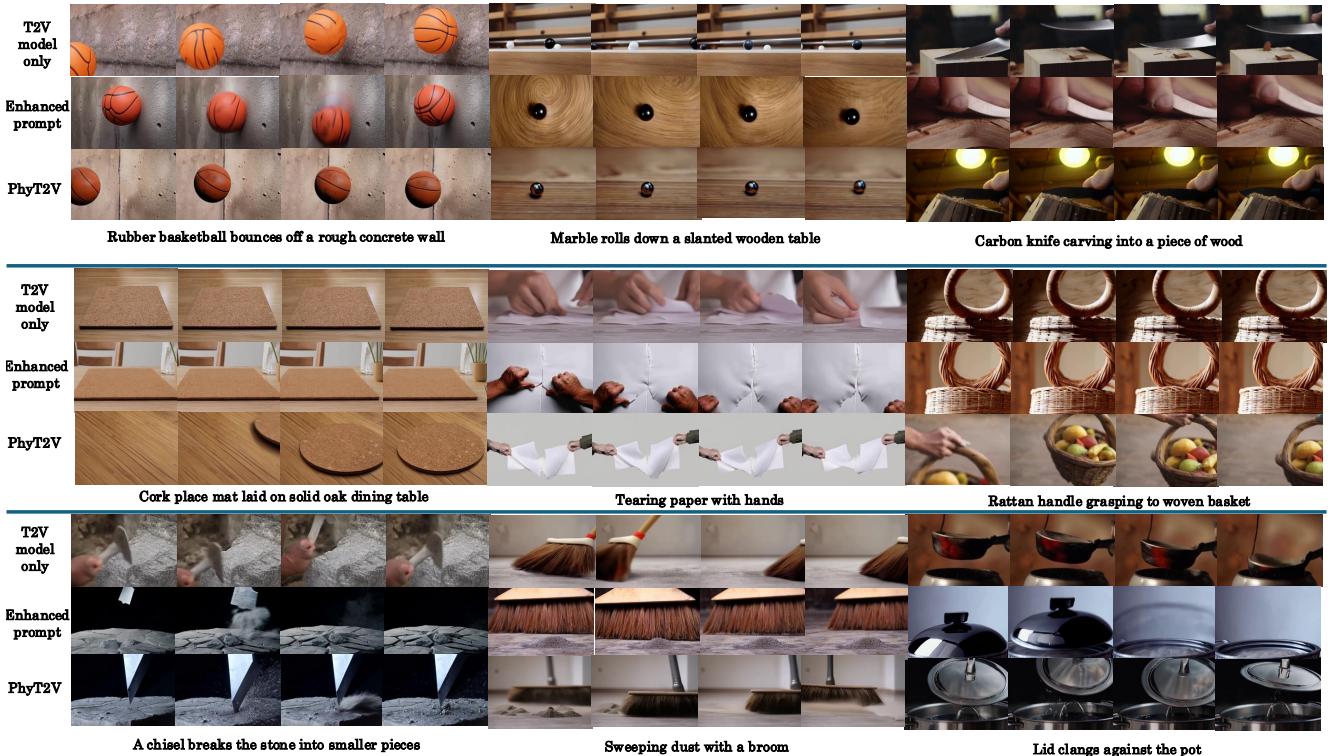




Figure 14. Video generation example on solid to solid specific prompt in VideoPhy dataset

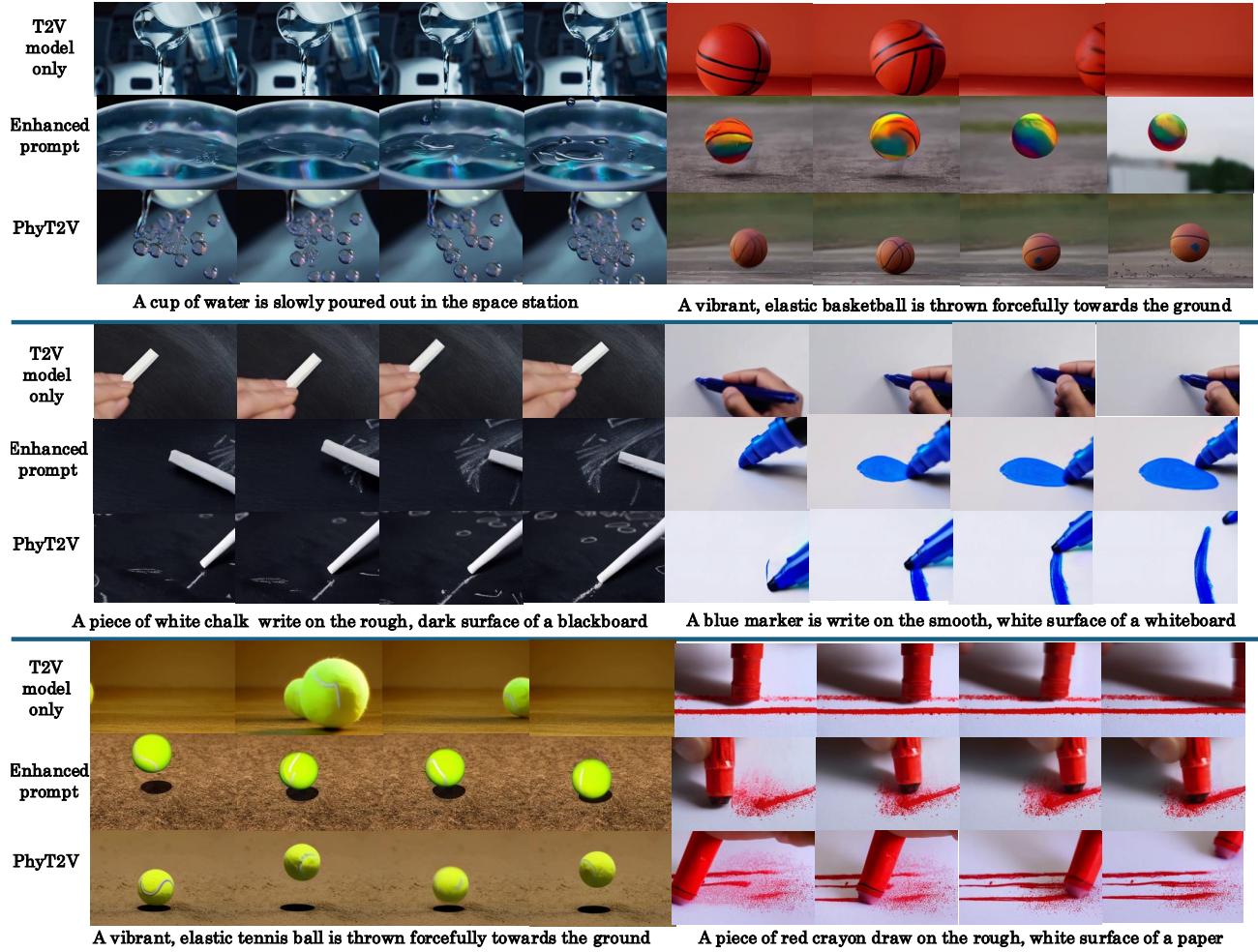


Figure 15. Video generation example on force specific prompt in PhyGenBench dataset

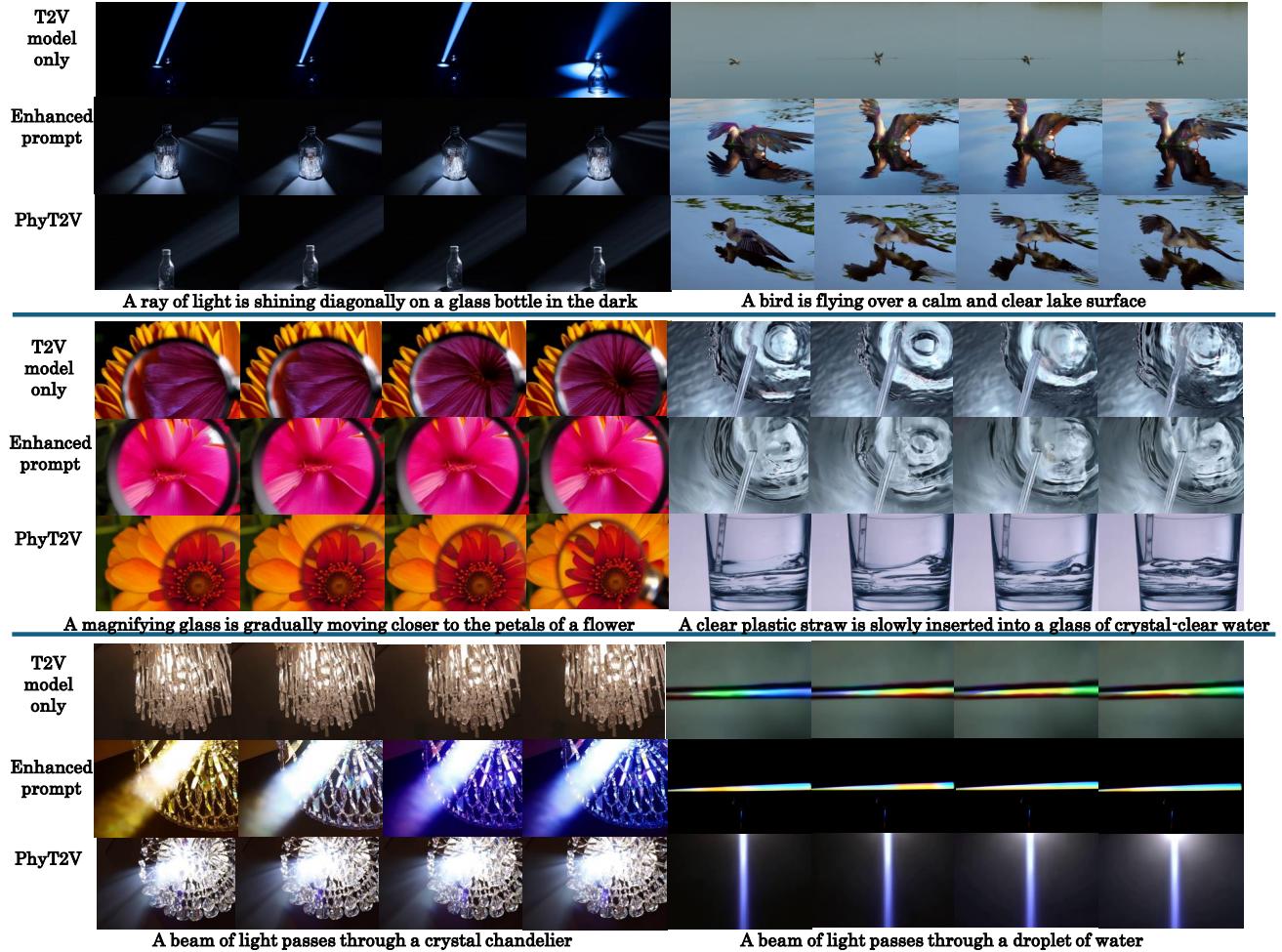


Figure 16. Video generation example on optics specific prompt in PhyGenBench

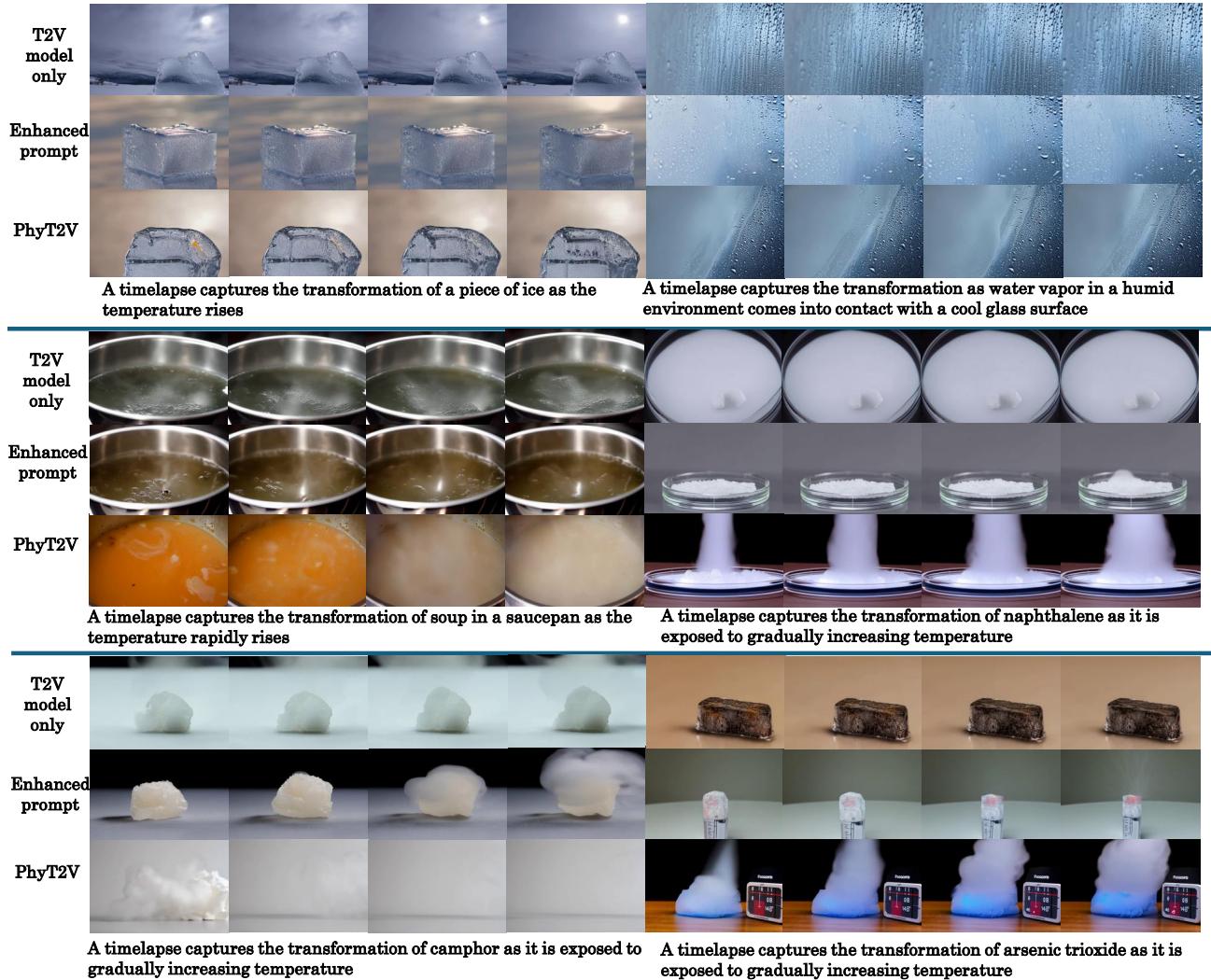
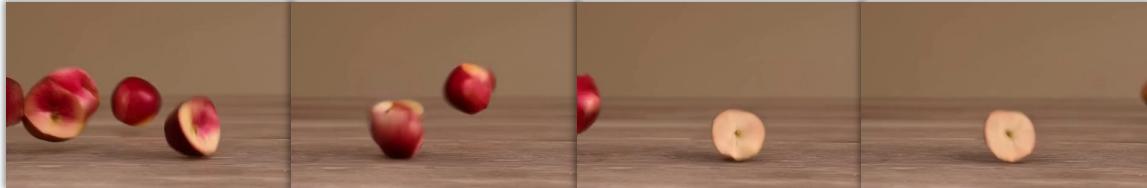


Figure 17. Video generation example on thermal specific prompt in PhyGenBench

Original prompt: An apple falls and bounces on the hard ground.



The key physical laws governing the apple's behaviour include Newton's Laws of Motion, the Law of Conservation of Energy, and the principles of partially elastic collisions and gravitational acceleration. 1. Gravitational Acceleration (Newton's Second Law of Motion): The apple falls because it is acted upon by the force of gravity, which causes it to accelerate towards the Earth's surface. According to Newton's Second Law, the force acting on the apple equals its mass times the gravitational acceleration - commonly around 9.8 m/s^2 on planet Earth. 2. Collision with the Ground (Partially Elastic and Inelastic Collisions): Upon hitting the hard ground, a collision occurs. The apple's largely inelastic nature means that a large portion of its energy gets absorbed in the form of heat and sound during collision, causing the apple not to bounce high. 3. Conservation of Energy: As the apple falls, it transforms potential energy into kinetic energy, with its velocity increasing continuously until impact. After bouncing, the apple follows the rules of projectile motion in gravity, accelerating upwards until it reaches a point where kinetic energy has been entirely converted back to potential energy - the point of zero velocity. The apple then starts falling again, continuing this process but with each bounce having less height due to energy losses in each collision.

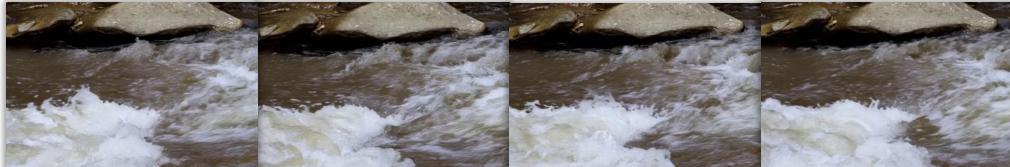
Mismatch: Apple vs. Apple Pieces: The user prompt describes an intact apple falling and bouncing on the ground. However, the video caption shows pieces of an apple falling and bouncing. This is a significant difference because the behaviour (including bouncing dynamics and deformation) of an intact apple and apple pieces upon impact would be expected to be quite different. Ground vs. Wooden Surface: The user prompt specifies the apple falls and bounces on 'hard ground.' In contrast, the video shows the apple pieces falling and bouncing on a 'wooden surface.' While we can assume both are 'hard,' we can't be sure they have the same hardness, which might affect the bouncing behaviour. Deformation of Apple: According to the user prompt, an intact apple is expected to fall and bounce without significant deformation. However, the video caption describes the apple pieces being compressed and crushed upon impact. This level of deformation is inconsistent with the user prompt description of a single apple bouncing.

Enhanced prompt: A ripe, bold red apple, defying its tree's pull, is seen in mid-air above a rough, hard terrain. The apple plummets, obeying gravity's irresistible command. Its descent accelerates, and it nears the unforgiving ground with increasing speed. Collision is inevitable. Upon impact, it momentarily squashes, absorbing a portion of its kinetic energy, before exhibiting a modest upward recoil. This bouncing trajectory, though significantly less than its initial fall, is agreed upon with the stringent laws of energy conservation. With each bounce, it ascends lesser, painting a perfect illustration of energy transformation and loss amid the persistent pull of gravity.



Figure 18. Refinement detail example on solid to solid specific prompt in VideoPhy dataset

Original prompt: A stone plunges into a fast-moving river, causing water to splash up.



The physics principles to note in this system include gravity, buoyancy, fluid resistance, and displacement of water by the falling stone. 1. Gravitational Acceleration (Newton's Second Law of Motion): The initial movement of the stone into the river is driven by gravity pulling on the stone's mass and creating acceleration. This leads to an initial downward motion of the stone. 2. Fluid Resistance: As the stone enters the water, it experiences fluid resistance which slows down its descent. This fluid resistance is proportional to the stone's velocity through the water and its size and shape. 3. Displacement of Water (Archimedes' Principle): As the stone plunges into the water, it displaces a volume of water equal to its own volume. According to the Archimedes' Principle, the buoyant force exerted on the stone is equal to the weight of the water displaced by the stone. The downward motion of the stone causes the water to be pushed aside and this sudden displacement causes the water to splash upwards. 4. Buoyancy: Depending on the mass and volume of the stone, buoyancy may play a role in determining whether the stone sinks or floats. A stone generally sinks because its density is greater than the water's density. However, if it is porous and air-filled, it could potentially float. The buoyant force working against gravity would determine the final resting position of the stone in the water.

Physical law:

Mismatch: Object Interaction with Water: The user prompt specifies a stone 'plunging' into a fast-moving river, implying that the stone is being dropped or thrown into the water from an external source, causing a splash. However, the video caption portrays a different scenario, where a large rock is already partially submerged in the water, around which the water flows and splashes. Absence of Splashing Caused by Stone: In line with the user prompt's narrative, the stone plunging into the water should cause a disturbance in the water, resulting in a visible splash. This crucial detail is missing in the video caption, which only mentions that water splashes because it encounters an obstacle - the rock - which is already partially submerged in it. Static vs Dynamic Position of the Stone: The user's prompt assumes the stone is actively interacting with the river by being dropped into it. It's a dynamic process. In the video caption, on the other hand, the rock is static, remaining stationary in the river while the river's water flows around it.

Enhanced prompt: A moderately-sized stone abruptly drops from the frame's upper edge onto a shimmering, fast-moving, and vigorous river, causing an impactful plunge. As it pierces the water's surface, the force causes a dramatic eruption of clear, sparkling droplets to violently splash upward. The displaced water ripples out in expanding concentric circles, effectively depicting the stone's powerful intrusion. The stone's downward trajectory continues unwavering, gradually obscured by the depth and swirling churn of the river. The re-arranged water reforms itself, concealing the stone beneath its ever-moving, turbulent surface, while the resulting splashes gradually fall back, a few rebellious droplets catching the light before surrendering to the relentless flow of the river.



Figure 19. Refinement detail example on solid to fluid specific prompt in VideoPhy dataset

Original prompt: Pouring milk into boiling tea.



The milk being poured into boiling tea should primarily follow the Law of Conservation of Mass and Energy, along with Thermodynamics and principles of fluid flow. 1. Conservation of Mass: As the milk is poured into the tea, the mass is conserved. This implies that the total mass of the milk and tea in the system remains constant during the process. 2. Thermodynamics (Heat Transfer): Due to the temperature difference between the milk and the boiling tea, heat will be transferred from the hotter object (the tea) to the cooler one (the milk). This heat transfer continues until both fluids reach a thermal equilibrium, where they have the same temperature. This process illustrates the Second Law of Thermodynamics, which states that heat will always move to areas of lower temperature until equilibrium is achieved. 3. Fluid Mechanics: The milk being poured is subjected to gravity and it flows via the principles of fluid dynamics, including viscosity and pressure. As per Pascal's Law, pressure applied to any part of a fluid transmits equally in all directions. Hence, the milk disperses throughout the tea. 4. Solubility: If viewed at the molecular level, the milk proteins and fats will disperse in the tea due to the natural molecular motion and interactions leading to the eventual homogeneous mixture of the two liquids. This highlights the principle of solubility and spontaneous mixing due to entropy increase.

Physical law:

Mismatch: The user prompt specifies pouring 'milk into boiling tea', but the video caption depicts 'milk being poured into a pot of simmering broth'. The substance into which the milk is being poured does not match the user's provided prompt. Incorrect Physical Phenomenon: The video refers to buoyancy, describing the milk as initially floating because 'it is less dense than the broth'. This doesn't align with the real-life action of pouring milk into tea, where the milk will generally blend with the tea quickly rather than floating. Severity of Heat: 'Boiling tea' implies a higher heat level than 'simmering broth', affecting the interaction between the milk and the tea (or broth) at different heat levels. The mismatch in describing temperature might lead to different outcomes in the two scenarios.

Enhanced prompt:

In a stoneware mug filled with scalding, steaming tea, cool, smooth milk cascades in a steady stream. As the dense white liquid merges with the boiling, amber tea, the color begins to shift, transforming into a richer, creamier hue. Gradually, any initial vestiges of division fade until the fluid becomes one harmonious blend, their temperatures equalizing in the warmth of the mug. Thermodynamics, fluid dynamics, and solubility principles are brilliantly on display, illustrating a soothing preparation ritual. The final shot reveals the comforting, perfectly steeped tea, gently wafting a fragrant steam, ready to sip.



Figure 20. Refinement detail example on fluid to fluid specific prompt in VideoPhy dataset

Original prompt: A cup of water is slowly poured out in the space station.



While in the space station, the cup of water would be subject to several important physical laws including the Law of Conservation of Mass, Newton's Laws of Motion, and the principle of surface tension. 1. Law of Conservation of Mass: Regardless of its location in space, the mass of the water remains the same. When the water is poured out from the cup, it doesn't disappear; rather it distributes in the area around it. 2. Newton's Laws of Motion: In zero gravity environment, actions and reactions will have a more noticeable impact. When water is poured from the cup, the act of forcing the water out would also push the cup in the opposite direction. This reaction is dictated by Newton's Third Law of Motion, which states that for every action, there is an equal and opposite reaction. 3. Surface Tension & Formation of Spheres: In space, without the influence of gravity, liquids naturally form a shape that gives the least surface area, which is a sphere. This is due to the cohesive forces between the molecules of the liquid (surface tension) which pulls the molecules together, thus forming a sphere. This is why when water is released into space, it forms globules or spheres that float around rather than spreading out like in earth's gravity. 4. Law of Conservation of Momentum: If the water is poured forcefully from the cup, the sum of the momentum of the cup and water before being poured (if in relative rest, it is zero) and after being poured will be conserved.

Physical law:

Absence of Zero Gravity Condition: The user prompt describes a cup of water being poured out in the space station, which would be an environment with negligible gravity – the 'zero-gravity' or microgravity environment. In microgravity, liquids like water form into spherical drops or z and float in place rather than producing a flowing stream downwards. However, the video caption describes the water being poured from a cup in a way that would only happen in a terrestrial setting with gravity: forming a stream and falling down with deformation due to gravity. Ignoring the Space-Station Setting: The user prompt specifies this action to be occurring in a space station. In contrast, the video caption makes no mention of the space station and erroneously depicts a possible laboratory setting implying gravity. Misrepresentation of Fluid Behavior in Microgravity: As there is virtually no gravity in the space station, the water would not flow out and fall as it does under the influence of gravity. It would float in the air as droplets or form larger blobs due to cohesive forces. Hence, the video's depiction of the water forming a stream and falling due to gravity conflicts with the user prompt's specification.

Mismatch:

Enhanced prompt: Amidst the ethereal tranquility of a space station, a cup gracefully tilts, gradually releasing an entrancing display of water. Unhindered by gravity, the water does not pour but emerges as perfect, shimmering spheres. Born of cohesion and surface tension, they form a mesmerizing constellation expanding from the cup's mouth. These watery orbs, imbued with spectral hues, disband from the cup and peacefully float off in the surrounding area, embodying freedom, echoing the whisper of zero gravity that holds sway here. Each sphere, a miniature world to itself, adrift and unhurried, is testament to the complexities yet serenity of life in a space-bound microgravity environment.



Figure 21. Refinement detail example on force specific prompt in PhyGenBench dataset

Original prompt: A bird is flying over a calm and clear lake surface, with the reflection of the bird appearing on the water surface.



The flight of the bird obeys various physical principles, notably Newton's Laws of Motion and the Law of Reflection while the reflection obeys the Law of Reflection and Refraction. 1. Bird Flight (Newton's Laws of Motion, Lift, and Drag Force): Three laws are particularly relevant. The first law of motion states that an object will remain at rest or in uniform motion unless acted upon by an external force. Therefore, bird must exert a force greater than or equal to its weight to maintain altitude and speed in flight. The third law of motion states that for every action, there is an equal and opposite reaction, which means that as a bird flaps its wings downward, a lifting force is generated that enables the bird to stay airborne. Lastly, the second law of motion, which states that the extent of acceleration depends on the object's mass and the amount of net force exerted on it, is applicable when the bird changes its flight speed or direction. Additionally, the bird must also overcome drag and gravity to maintain forward momentum. 2. Reflection of Bird (Law of Reflection): The light that enables us to see the bird's reflection comes from the sun, hitting the bird and then reflecting off the lake to our eyes. According to the Law of Reflection, the angle at which the light hits the bird (the angle of incidence) is equal to the angle at which it bounces off the lake's surface (the angle of reflection). 3. Distortion in Reflection (Refraction): Though not explicitly stated, we can infer that any distortions in the bird's reflection could be due to the physical property of water called refraction. Light speed changes when it moves from the air to water, causing the light to bend and hence the image to distort.

Physical law:

Mismatch:
Position and Motion of the Bird: The user prompt describes a celestial bird sailing across a tranquil lake, creating a moving reflection on the water surface as it modifies its flight path. However, the video caption depicts a bird standing stationary on the water surface, flapping its wings but not flying or sailing. Reflection on the Water Surface: The user prompt refers to an enchanting double image of the bird and its flight reflected on the mirror-like lake. On the contrary, in the video, the reflected image correlates only to the bird's stationary position and wing movements. Interplay of Light: The prompt speaks to a captivating light play, with sunlight glancing off the bird's plumage and the water's surface. This aspect is completely unaddressed in the video caption. Effects of Wind: The user prompt mentions a soft wind causing gentle ripples on the water's surface, distorting the bird's reflection. In contrast, the ripples in the video caption occur solely due to the bird's wing movement. Physics Misinterpretations: The video caption erroneously attributes the bird's ability to stand on water to surface tension, while in reality, most birds float on water due to their lightweight bodies and waterproof feathers. The physics laws of surface tension aren't accurately portrayed.

Enhanced prompt:
A celestial bird, feathered in iridescent plumage, effortlessly traces a majestic path across a tranquil and invitingly clear lake surface. The bird creates an enchanting double image, reflected with verisimilitude on the mirror-like water, revealing every intricacy of its sublime flight. The reflection of sunlight glimmers off the glossy plumage of the bird and the lustrous water surface, illuminating the scene with a captivating interplay of light. As the bird banks and alters its trajectory, the mirrored image dances and ripples with each movement, distorting tenderly under the soft breeze that ruffles the serene lake's surface.



Figure 22. Refinement detail example on optics specific prompt in PhyGenBench dataset

Original prompt: A timelapse captures the transformation of arsenic trioxide as it is exposed to gradually increasing temperature.



The key physical laws that arsenic trioxide obeys in this scenario are primarily focused within the realms of thermodynamics and phase transition. 1. Heat Transfer and Thermodynamics: As the arsenic trioxide is exposed to increasing temperature, it's subject to the principles of thermodynamics, specifically the first law of thermodynamics which essentially states that energy cannot be created or destroyed, but only transformed. In this case, the thermal energy is transferred from the surrounding area to the arsenic trioxide, causing the chemical to heat up.

Physical law: 2. Phase Transition: Arsenic trioxide should experience phase transition under sufficient heat. This is described by the phase transition theory, a subset of thermodynamics. Initially, arsenic trioxide at room conditions is in solid form, and as the thermodynamic equilibrium changes with heating, it undergoes a phase transition process, changing from solid to liquid, and eventually to gas, each phase bound by a certain range of temperature and pressure. The transformations are not instantaneous but happen over a period of time, thereby can be captured in a timelapse video. 3. Expansion due to Heating: The arsenic trioxide may also physically expand as it heats, which is explained by the principle of thermal expansion. As arsenic trioxide heats up, its molecules move more rapidly, causing the substance to expand. However, this may be more apparent in larger masses of matter and may not show significant effect in a small amount of arsenic trioxide. 4. Gas Laws: If the arsenic trioxide reaches a gaseous state, it then follows the gas laws, including Charles's Law and Boyle's Law, which describe the relationships between the volume, pressure, and temperature of gases.

Mismatch: Absence of Transformation and Temperature Change: The user prompt asks for a video illustrating the change in arsenic trioxide when exposed to a gradually increasing temperature starting from room temperature. In contrast, the video caption describes a static close-up shot of arsenic trioxide without any transformation or interaction with a varying temperature environment. Misplaced Emphasis on Camera Physics: The prompt requests a physics demonstration or experiment involving chemical transformation under temperature changes. However, the video caption instead focuses on the physics related to the camera's zoom and focus, such as light refraction and lens adjustments. These aspects, although they involve physics, are completely unrelated to the user prompt regarding the thermochemical behavior of arsenic trioxide, indicating a significant mismatch. Misinterpretation of User Prompt: The video caption does not address the desired timelapse showing how arsenic trioxide changes when the temperature rises from room temperature. Instead, it provides a static shot of the substance without transformation or interaction with temperature changes. This deviates from the user's request, which involves observation of physical changes under different thermal conditions.

Enhanced prompt: A timelapse illustrates the transition of a crystalline lump of arsenic trioxide going through a radiant transformation. As the temperature gradually rises from room temperature, captured by a subtly placed thermometer, the seemingly motionless arsenic trioxide begins to stir. Initially, microscopic tremors agitate the lump as it slowly warms, and increasingly visible motions ensue as the temperature rises further. Solid arsenic oxide soon begins to liquefy, with sparks capturing the process. Finally, the gas begins to wisps upwards. The entire spectacle reflects the grandeur of thermodynamics in action.



Figure 23. Refinement detail example on thermal specific prompt in PhyGenBench dataset

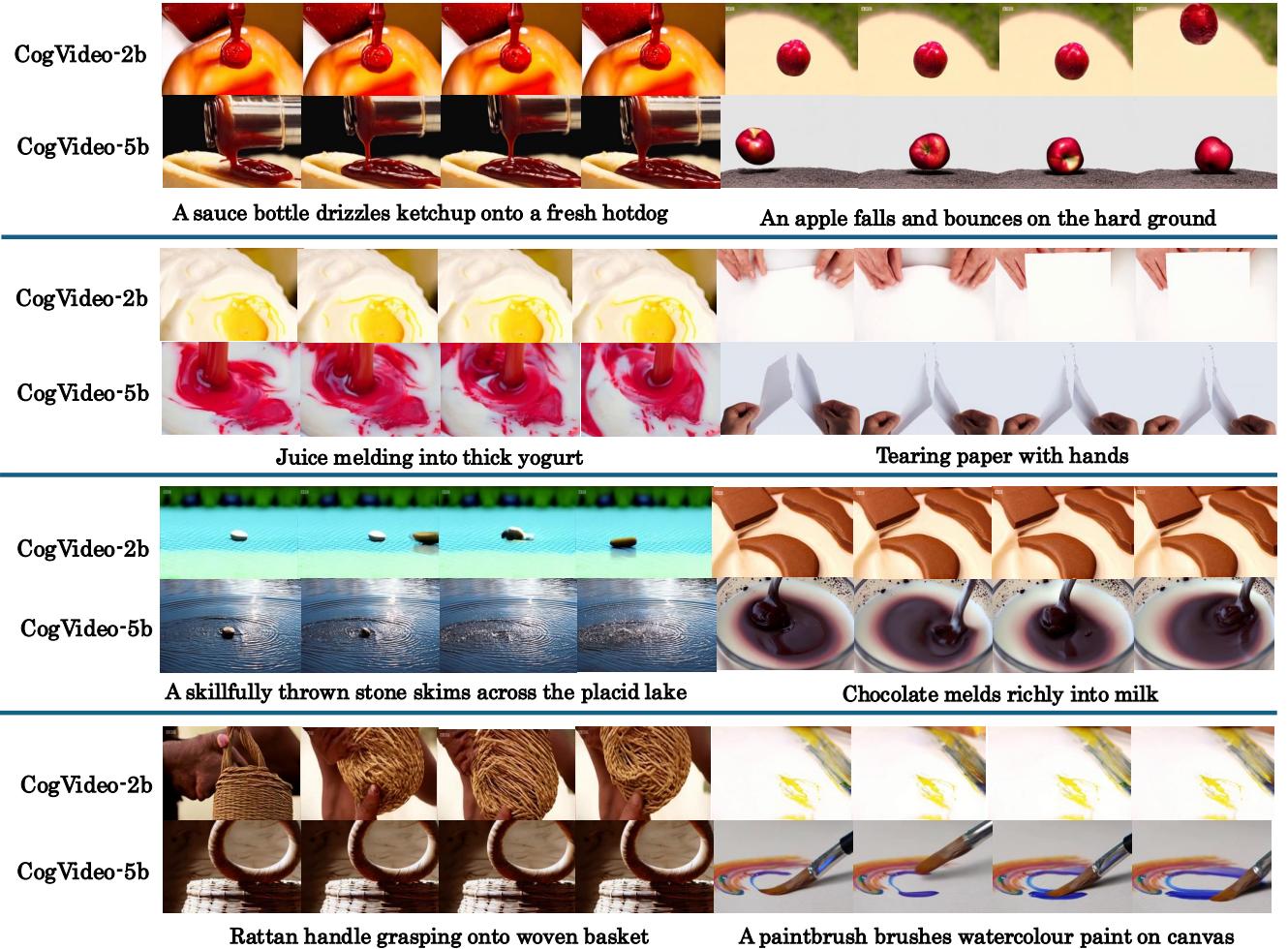


Figure 24. Ablation study on model size

Prompt

Instruction

You are a physics expert. Your task is to identify the main object in the given user prompt and provide the physical laws in reality the main object should obey with as much detail as possible in a descriptive way without giving formulas. Some in-context examples are provided for your reference, and you need to finish the current task.

In-context examples

User prompt: a rubber ball hits the ground and then bounces up

Main object: rubber ball,

Physical law: The primary physical law that should be obeyed by the video is Newton's Law of Motion along with the Law of Conservation of Energy, particularly focusing on elastic collisions and gravitational acceleration. 1. Gravitational Acceleration (Newton's Second Law of Motion): As the rubber ball falls toward the ground, it is acted upon by the force of gravity. According to Newton's Second Law, the force acting on the ball is the product of its mass and the gravitational acceleration, typically 9.8 m/s^2 near the surface of the Earth. 2. Collision with the Ground (Elastic and Inelastic Collisions): When the ball hits the ground, a collision occurs. In reality, rubber balls exhibit partially elastic behavior, meaning that some energy is lost to heat and deformation during the impact. This leads to a bounce with less energy than the initial fall, and the ball does not reach the original height from which it was dropped. 3. Conservation of Energy: As the ball falls, its potential energy is converted into kinetic energy, the velocity of the ball keeps increasing until the impact. Once the ball bounces back up, it follows the rules of projectile motion under gravity, accelerating upwards until the velocity reaches zero at its highest point, where all kinetic energy has been converted back to potential energy. The ball then begins its downward motion again, repeating the cycle but with diminishing height due to energy loss at each bounce.

Current task

User prompt: An apple falls and bounces on the hard ground.

Let's think step by step.

Output

Main object: apple

Physical law:

1. Gravitational Acceleration

As the apple falls, it is influenced by the force of gravity. Gravity pulls the apple downward, causing it to accelerate toward the Earth. The apple's acceleration due to gravity is constant near the Earth's surface, regardless of its mass. Its velocity increases steadily during free fall unless influenced by air resistance.

2. Air Resistance (Drag Force)

While the apple is falling, it encounters air resistance, a force opposing its motion. The magnitude of air resistance depends on the apple's shape, size, and speed, as well as the density of the air. If the apple falls from a sufficient height, it may reach terminal velocity—a constant speed where the downward force of gravity equals the upward drag force.

3. Impact with the Hard Ground

When the apple hits the ground, it undergoes a collision. This interaction is governed by:

Newton's Third Law of Motion: The apple exerts a force on the ground, and the ground exerts an equal and opposite force on the apple.

Impulse and Momentum: During the collision, the apple's momentum changes rapidly over a short period, depending on its velocity and mass.

Deformation and Elasticity: The apple, being a soft object, deforms on impact. The ground, being hard, undergoes minimal deformation. The extent to which the apple regains its shape determines the height of its bounce.

4. Conservation of Energy

Before the fall, the apple has gravitational potential energy due to its height above the ground.

As it falls, this potential energy is converted into kinetic energy.

At the moment of impact, some of the kinetic energy transforms into other forms:

Elastic Potential Energy (as the apple compresses and deforms).

Thermal Energy (heat generated due to internal friction in the apple and at the contact surface).

Sound Energy (the noise created upon impact).

During the bounce, the elastic potential energy in the apple partially converts back into kinetic energy. However, some energy is lost, resulting in a lower bounce height.

5. Coefficient of Restitution (COR)

The apple's ability to bounce depends on the COR, a measure of how elastic the collision is. For the apple-ground system, the COR is less than 1, indicating an inelastic collision where energy is lost to deformation and other processes.

6. Post-Bounce Motion

After bouncing, the apple follows a parabolic trajectory dictated by gravitational forces. Its upward motion decelerates due to gravity until it momentarily stops at its maximum height. If the apple bounces multiple times, each subsequent bounce will be shorter due to cumulative energy losses.

Figure 25. The original prompt and output

Prompt

Instruction

Your task is to identify the main object in the given user prompt and provide the physical laws in reality the main object should obey with as much detail as possible in a descriptive way without giving formulas. Some in-context examples are provided for your reference, and you need to finish the current task.

In-context examples

User prompt: a rubber ball hits the ground and then bounces up

Main object: rubber ball,

Physical law: The primary physical law that should be obeyed by the video is Newton's Law of Motion along with the Law of Conservation of Energy, particularly focusing on elastic collisions and gravitational acceleration. 1. Gravitational Acceleration (Newton's Second Law of Motion): As the rubber ball falls toward the ground, it is acted upon by the force of gravity. According to Newton's Second Law, the force acting on the ball is the product of its mass and the gravitational acceleration, typically 9.8 m/s^2 near the surface of the Earth. 2. Collision with the Hard Ground (Inelastic Collision): When the ball hits the ground, a collision occurs. In reality, rubber balls exhibit partially elastic behavior, meaning that some energy is lost to heat and deformation during the impact. This leads to a bounce with less energy than the initial fall, and the ball does not reach the original height from which it was dropped. 3. Conservation of Energy: As the ball falls, its potential energy is converted into kinetic energy, the velocity of the ball keeps increasing until the impact. Once the ball bounces back up, it follows the rules of projectile motion under gravity, accelerating upwards until the velocity reaches zero at its highest point, where all kinetic energy has been converted back to potential energy. The ball then begins its downward motion again, repeating the cycle but with diminishing height due to energy loss at each bounce.

Current task

User prompt: An apple falls and bounces on the hard ground.

Let's think step by step.

Output

Main object: apple

Physical law:

1. Gravitational Acceleration (Newton's Second Law of Motion):

As the apple falls toward the ground, it is acted upon by the force of gravity. This force causes the apple to accelerate downward at a constant rate, typically 9.8 m/s^2 near the Earth's surface. The apple's velocity increases steadily as it descends due to this constant acceleration.

2. Collision with the Hard Ground (Inelastic Collision):

When the apple hits the hard ground, a collision occurs. Unlike a rubber ball, the apple is less elastic and more prone to deformation upon impact. The collision is primarily inelastic, meaning a significant portion of kinetic energy is lost to internal deformation, heat, and sound. This results in a less pronounced bounce compared to more elastic objects.

3. Conservation of Energy:

During the fall, the apple's gravitational potential energy is converted into kinetic energy, increasing its speed until impact. Upon collision with the ground, some kinetic energy transforms into other forms of energy: Heat Energy: Generated due to internal friction within the apple's structure. Sound Energy: Produced from the noise of the impact. Energy of Deformation: Used in changing the shape of the apple upon impact. If the apple bounces, the remaining kinetic energy propels it upward, but it will not reach the original height due to these energy losses.

4. Newton's Third Law of Motion:

During the collision, the apple exerts a force on the ground. The ground exerts an equal and opposite force back on the apple. This reaction force can cause the apple to bounce back upward, depending on how much kinetic energy remains after the impact.

5. Air Resistance:

As the apple falls and potentially bounces upward, it experiences air resistance. Air resistance opposes the motion of the apple through the air, reducing its acceleration during the fall. During the upward motion after a bounce, air resistance works against the apple's ascent, causing it to slow down more quickly.

6. Material Properties and Deformation:

The apple's ability to bounce is influenced by its material properties, such as elasticity and structural integrity. Being less elastic, the apple absorbs more energy through deformation upon impact. This absorption of energy reduces the amount of kinetic energy available to propel the apple back upward, resulting in a lower bounce height compared to elastic objects like rubber balls.

7. Energy Dissipation:

Energy lost during the collision and subsequent motions ensures the apple will not bounce back to its original height. Over successive bounces (if any), the apple's kinetic energy continues to decrease due to ongoing energy losses. Eventually, the apple comes to rest as all its kinetic energy is dissipated through deformation, heat, sound, and overcoming air resistance.

Figure 26. The prompt without the role indicator and output

Instruction
 You are a physics expert. Your task is to identify the main object in the given user prompt and provide the physical laws in reality the main object should obey with as much detail as possible in a descriptive way without giving formulas.
Prompt
 User prompt: An apple falls and bounces on the hard ground.
 Let's think step by step.
Main Object:
 The main object in this scenario is the apple.
Physical Laws Governing the Apple's Motion
 Let us analyze the motion of the apple step by step:
 Step 1: Initial Fall (Gravitational Pull)
 The apple starts its motion due to gravity, which exerts a constant downward force. Gravity pulls the apple toward the Earth, causing it to accelerate uniformly. This is governed by Newton's Second Law of Motion, where the net force on the apple results in an acceleration proportional to its mass.
 Step 2: Interaction with Air (Air Resistance)
 As the apple falls, it moves through the air and experiences air resistance, a force that opposes its downward motion. Air resistance depends on factors like the apple's speed, surface area, and shape, as well as the density of the air. If the apple falls from a sufficient height, it could reach a state called terminal velocity, where the upward force of air resistance balances the downward force of gravity, causing it to fall at a constant speed.
 Step 3: Impact with the Ground (Collision Dynamics)
 When the apple strikes the hard ground, it undergoes a collision. The interaction involves several physical principles: Newton's Third Law of Motion: The apple exerts a force on the ground, and the ground exerts an equal and opposite force on the apple. Momentum Transfer: The apple's momentum changes rapidly due to the impact. The exact nature of this change depends on the time duration of the collision and the material properties of the apple and the ground. Deformation and Elasticity: The apple deforms upon impact due to its soft and elastic nature. The hard ground, being rigid, undergoes negligible deformation.
 Step 4: Energy Transformation at Impact
 As the apple falls, its gravitational potential energy is converted into kinetic energy. Upon hitting the ground, part of the kinetic energy is transformed into: Elastic Potential Energy: Stored as the apple compresses during the collision. Thermal Energy: Generated through internal friction within the apple and at the contact point. Sound Energy: Produced as the collision generates noise. Energy losses during the impact mean that the apple does not fully regain its initial potential energy.
 Step 5: The Bounce (Coefficient of Restitution)
 The apple rebounds after the impact, driven by the release of elastic potential energy stored during its deformation. The height of the bounce is determined by the coefficient of restitution (COR), which measures the elasticity of the collision. For an apple-ground interaction, the COR is less than 1, indicating that the collision is inelastic, with significant energy losses. Consequently, the apple bounces to a height lower than its original drop point.
 Step 6: Post-Bounce Motion
 After the bounce, the apple follows a parabolic trajectory determined by gravity. Its upward motion slows as it loses kinetic energy and eventually stops momentarily at its peak height. The apple then falls back under the influence of gravity. If subsequent bounces occur, each is progressively shorter due to cumulative energy losses to heat, sound, and internal deformation. The apple eventually comes to rest due to friction and dissipation of energy.
Relevant Physical Laws
 1. Newton's Laws of Motion. First Law: The apple remains at rest or in motion unless acted upon by a force (gravity, air resistance, ground collision). Second Law: The net force on the apple results in its acceleration toward the Earth. Third Law: During the collision, the ground exerts an equal and opposite force on the apple.
 2. Law of Universal Gravitation. The gravitational force between the Earth and the apple governs its downward motion.
 3. Energy Conservation and Transformation. Energy is converted between potential energy, kinetic energy, and other forms (elastic potential, thermal, sound). Energy losses due to inelasticity reduce the height of subsequent bounces.
 4. Air Resistance Principles. Drag forces oppose the apple's motion, particularly during free fall.
 5. Collision Dynamics and Elasticity. The behavior of the apple during and after impact is influenced by the deformation, energy losses, and the coefficient of restitution. By analyzing the motion in these steps, the apple's fall, impact, and bounce can be fully described.

Figure 27. The prompt without the in-context example and output

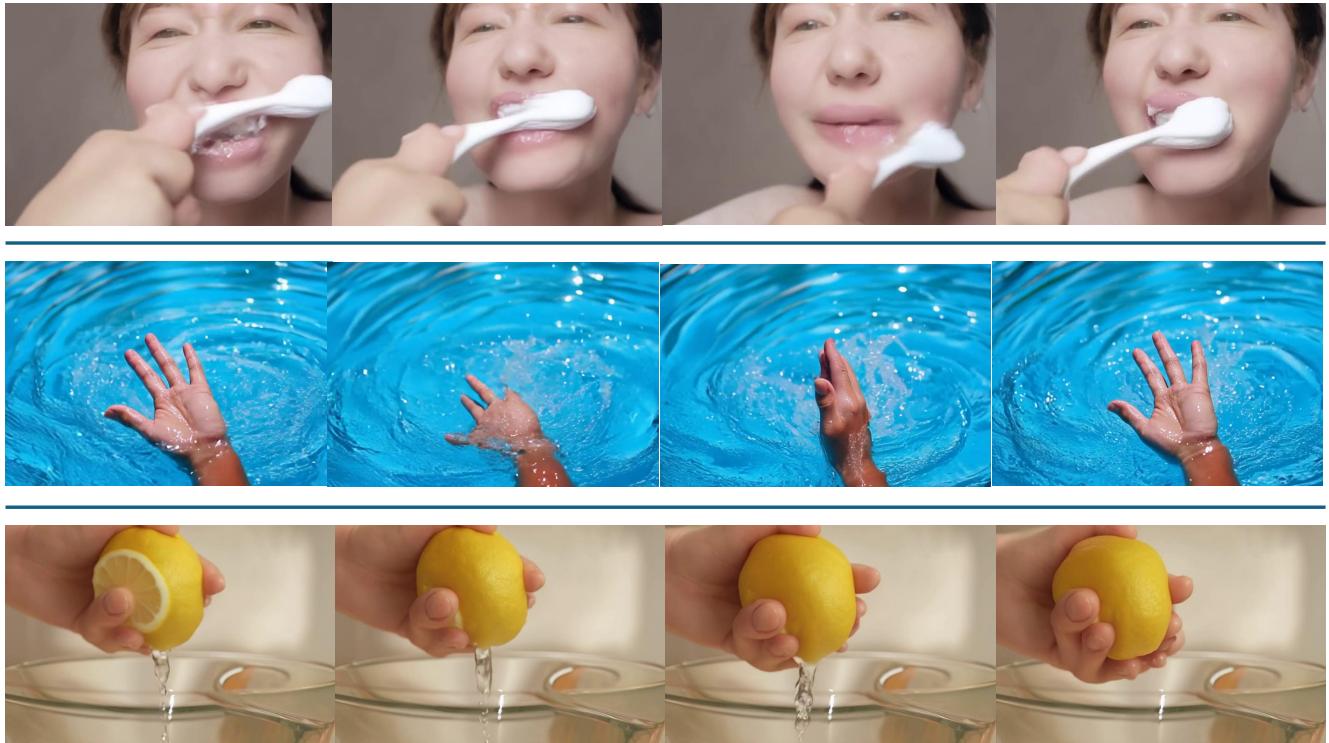


Figure 28. Failure case example on human hand related video generation