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

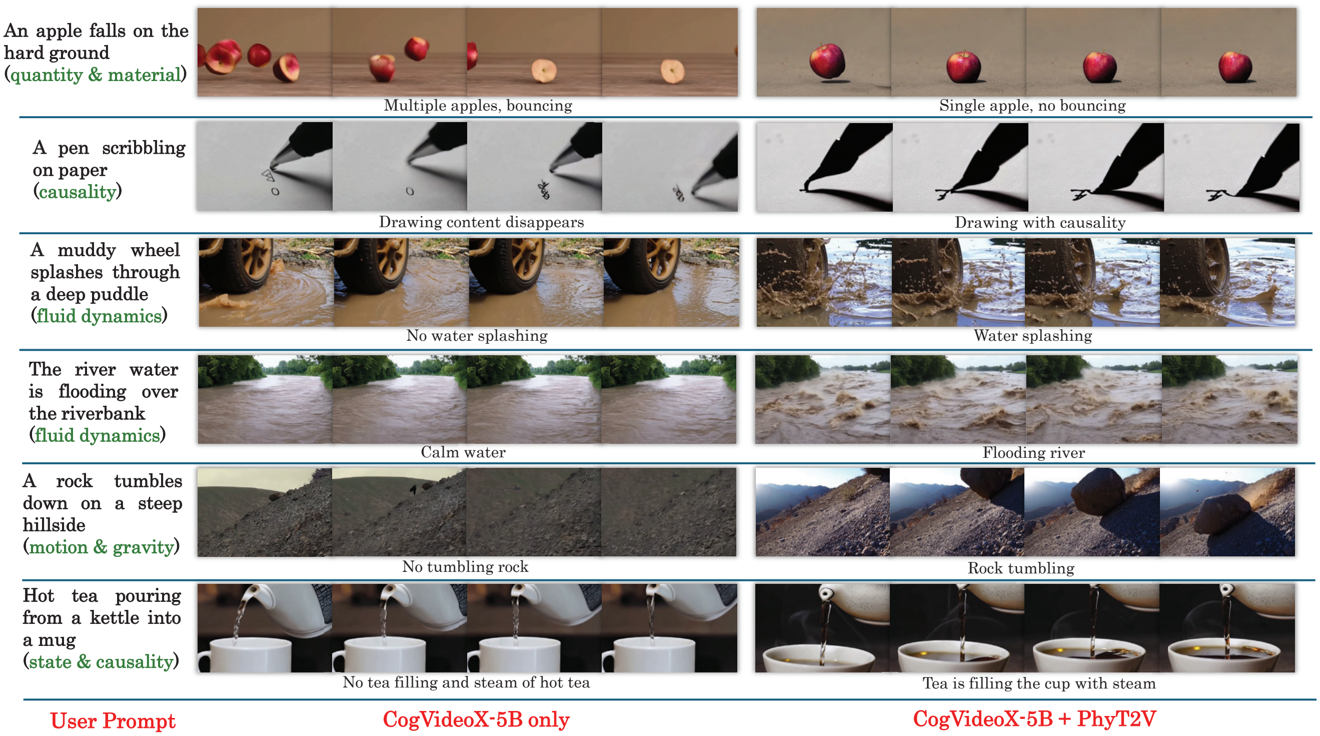

Figure 1. *Left*: videos generated by the current text-to-video generation model (CogVideoX-5B [50]) cannot adhere to the real-world physical rules (described in brackets following the user prompt). *Right*: our method PhyT2V, when applied to the same model, better reflects the real-world physical knowledge.

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 **Phy**scial-grounded **T**ext-**to**-**V**ideo (**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

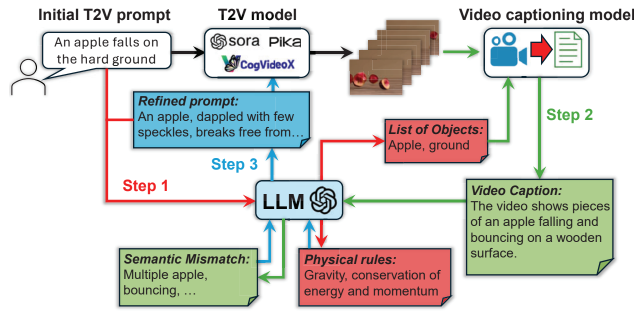

Figure 2. One iteration of video and prompt self-refinement in PhyT2V. Such self-refinement will be iteratively conducted in multiple rounds until the quality of generated video is satisfactory.

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 physcial-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 disconnected 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].

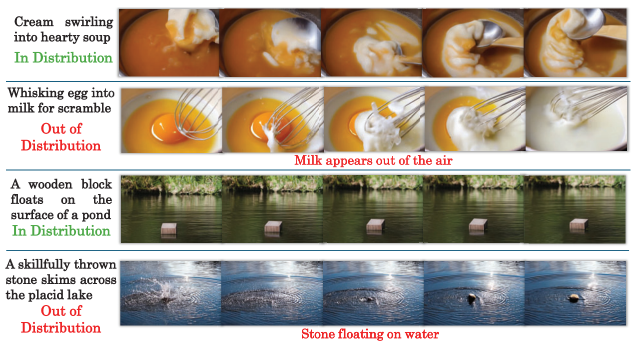

Figure 3. Examples of videos generated from in-distribution and out-of-distribution prompts, using the CogVideoX-5B model

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 data[1]. 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.

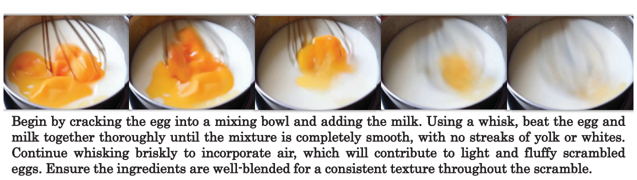

Figure 4. A video generated by enhancing the out-of-distribution prompt "Whisking egg into milk for scramble" in Figure 3

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 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.

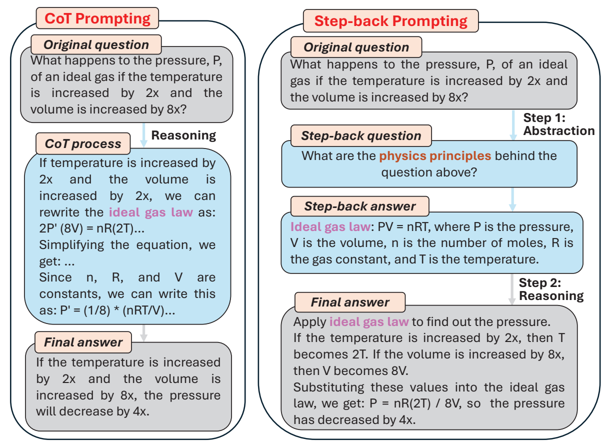

Figure 5. Examples of CoT and step-back reasoning

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,

[1]In 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.

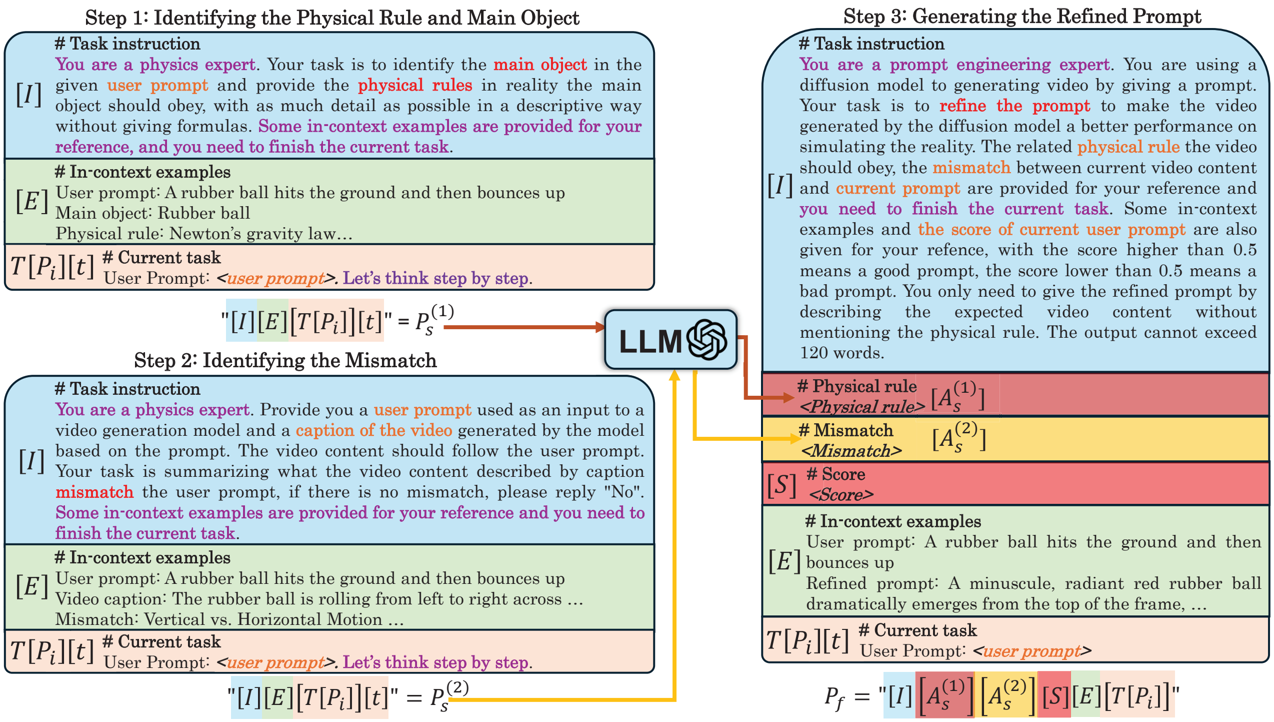

Figure 6. Our design of PhyT2V, illustrated by one round of video refinement consisting of three steps. Texts in brown are inputs from previous step. Texts in red are outputs to the next step; Texts in purple are prompts to trigger LLM reasoning

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., $p' = f_{\text{enhance}}(p, f_{mismatch}(C(V(p)), p), f_{phy}(p), \theta)$, 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 video[2]. 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.

[2]This 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 dataset, the binary outputs from all prompts are averaged.

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

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	0.21
	SA	0.41	0.32
Promptist [21]	PC	0.25	0.19
	SA	0.39	0.33

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	0.20
	SA	0.23	0.23
Promptist [21]	PC	0.32	0.19
	SA	0.24	0.21

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.

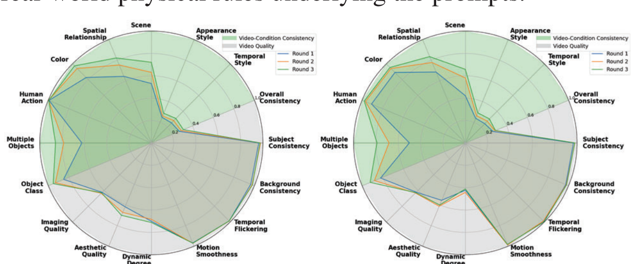

Figure 7. PhyT2V VBench evaluation results with CogVideoX-5B (left) and OpenSora (right)

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 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.

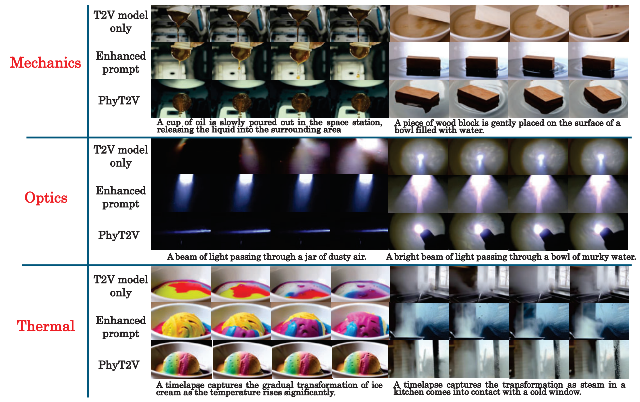

Figure 8. Examples of videos generated using different categories of prompts in the PhyGenBench dataset

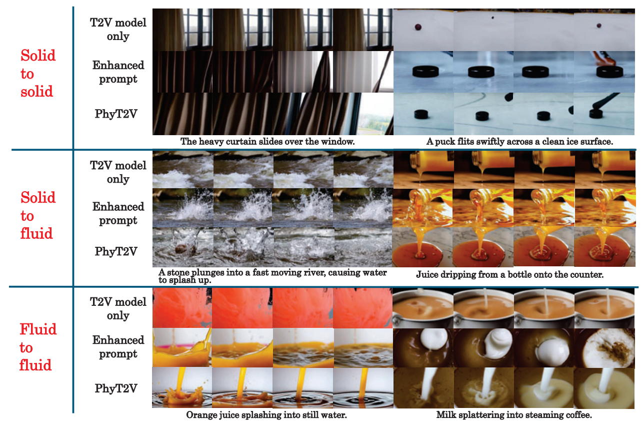

Figure 9. Examples of videos generated using different categories of prompts in the VideoPhy dataset

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.

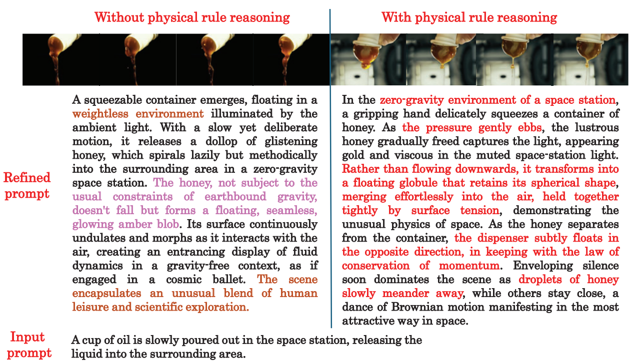

Figure 10. Ablation study on Step 1 of physical rule reasoning

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

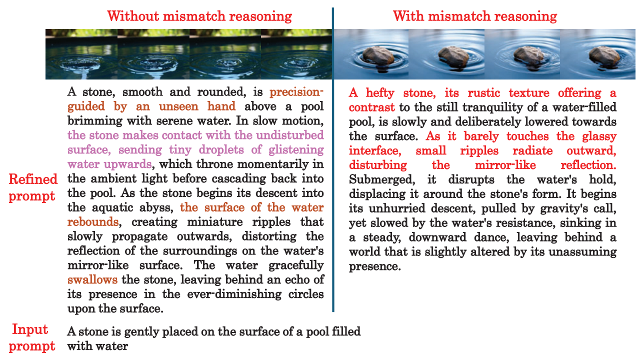

Figure 11. Ablation study on Step 2 of mismatch reasoning

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 insightfulcomments 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 realtime 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, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 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 Goncearenco, 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, Yunsen 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 Yuexian 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, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan 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, 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. 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