PhyDetEx: Detecting and Explaining the Physical Plausibility of T2V Models

Zeqing Wang[1,3], Keze Wang[1†], and Lei Zhang[2,3†]

[1] Sun Yat-sen University, China
[2] Hong Kong Polytechnic University, Hong Kong SAR, China
[3] OPPO Research Institute, China

Abstract. Driven by the growing capacity and training scale, Text-to-Video (T2V) generation models have recently achieved substantial progress in video quality, length, and instruction-following capability. However, whether these models can understand physics and generate physically plausible videos remains a question. While Vision-Language Models (VLMs) have been widely used as general-purpose evaluators in various applications, they struggle to identify the physically impossible content from generated videos. To investigate this issue, we construct a **PID** (**P**hysical **I**mplausibility **D**etection) dataset, which consists of a *test split* of 500 manually annotated videos and a *train split* of 2,588 paired videos, where each implausible video is generated by carefully rewriting the caption of its corresponding real-world video to induce T2V models producing physically implausible content. With the constructed dataset, we introduce a lightweight fine-tuning approach, enabling VLMs to not only detect physically implausible events but also generate textual explanations on the violated physical principles. Taking the fine-tuned VLM as a physical plausibility detector and explainer, namely **PhyDetEx**, we benchmark a series of state-of-the-art T2V models to assess their adherence to physical laws. Our findings show that although recent T2V models have made notable progress toward generating physically plausible content, understanding and adhering to physical laws remains a challenging issue, especially for open-source models. Our dataset, training code, and checkpoints are available at https://github.com/Zeqing-Wang/PhyDetEx.

Keywords: Text-to-Video Generation · Physical Plausibility Detection · Vision-Language Models

1 Introduction

Recent advances in Text-to-Video (T2V) generation models have significantly improved the visual quality, temporal consistency, *etc.*, of the generated videos [17, 25,26,35,48,60], partly building upon the rapid progress of high-quality synthetic data annotation [1,14,32,56] and reinforcement learning algorithms for diffusion

† Corresponding author

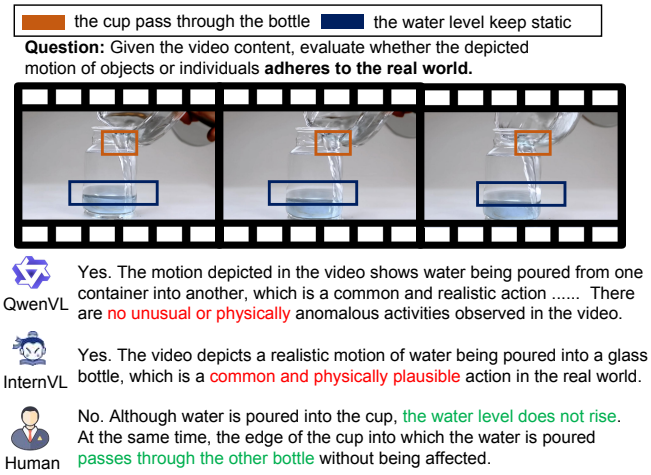

Fig. 1: Illustration of the physical implausibility detection task. Given a video where water is poured into a bottle, but the water level remains static and the cup passes through the bottle, humans can easily identify the violation of physical laws. However, current powerful VLMs (e.g., QwenVL and InternVL) incorrectly judge the motion as physically plausible, highlighting their difficulty in recognising implausible dynamics.

models [46]. With the rapid scaling of model capacity and multimodal training data, these models are increasingly viewed as promising candidates toward realising a world model [21,24,50], which can describe the dynamics and causality of the physical world. However, despite impressive progress, existing T2V models often produce physically implausible content that contradicts basic real-world principles [5,6,54]. Detecting and explaining these violations is crucial to measure how well current T2V models can understand the physical world.

Several studies attempt to assess the physical plausibility of generated videos using either pre-defined prompts [5,6,37,59] or detectors trained with extensive human annotations [5,6,53]. However, the first kind of methods are inherently restricted by their reliance on carefully designed prompts, while the second suffer from poor transferability and limited interpretability, which often produce only coarse plausibility scores without revealing why physical laws are violated. Consequently, these approaches remain narrow in scope and generality, underscoring the need for a more generalizable and interpretable evaluation framework that can handle a wide range of generated videos.

A natural approach to evaluate the physical plausibility of the generated videos with an explanation is to leverage Vision-Language Models (VLMs), which have achieved remarkable success across diverse perception and reasoning tasks [3,9,10,12,33]. VLMs are trained on massive visual–text corpora [13] and possess strong generalization and interpretability through textual reasoning [23,64]. It is intuitive to use them as general evaluators to identify physically implausible content in videos. However, recent evidence shows that even the most capable VLMs struggle to recognise physically implausible content gen-

erated by T2V models [4, 54]. As shown in Fig. 1, VLMs often fail to detect implausible content that is easy for humans to perceive. This motivates us to investigate whether we can endow VLMs, which have shown strong capability in many downstream visual tasks, with the ability to detect and explain the physical implausibility in generated videos.

To answer the above question, we first conduct a preliminary experiment by prompting several state-of-the-art VLMs to detect the physical plausibility in videos with varying degrees of prior information, without any additional training. We observe that their detection accuracy changes substantially depending on the given priors, suggesting that VLMs may possess certain physical knowledge, yet they struggle to apply this knowledge to unseen implausible cases without proper training. Therefore, we first construct a **P**hysical **I**mplausibility **D**etection (**PID**) dataset, which includes a *train split* and a *test split*. The **train split** consists of 2,588 paired videos, where each pair consists of a physically plausible (positive) video and an implausible (negative) counterpart. The implausible videos are generated by carefully rewriting captions to induce T2V models to produce physically impossible videos. The **test split** consists of 500 videos, including 250 physically implausible videos with manually annotated content of physical violations and 250 physically plausible counterparts. The implausible videos are collected from various T2V models that unintentionally generate abnormal content under normal prompts, while the plausible ones are drawn from both real-world videos and high-quality T2V generations.

Using the PID train split, we propose a lightweight fine-tuning approach that requires only minimal computational cost (via LoRA-based parameter-efficient tuning). The resulting model, namely **PhyDetEx** (**Phy**sical Plausibility **Det**ector and **Ex**plainer), empowers VLMs to (i) accurately identify physically implausible events and (ii) produce textual explanations of the violated physical principles. Our experiments demonstrate that PhyDetEx achieves significant improvements in detecting physical implausibility, validating that VLMs indeed possess an implicit understanding of physics that can be unlocked with a small number of contrastive examples. Finally, we apply PhyDetEx as an evaluation tool to benchmark the physical plausibility of current T2V models. Our results reveal that while recent closed-source models have made remarkable strides toward generating physically consistent videos, understanding and adhering to physical laws remain a substantial challenge, especially for open-source models. These findings highlight both the progress and the limitations of current T2V systems in genuine world modeling.

In summary, our contributions are as follows. First, we introduce **PID**, a comprehensive benchmark for physical implausibility detection of T2V models, which includes a *test split* of 500 annotated videos (250 physically plausible and 250 implausible) and a *train split* of 2,588 plausible/implausible paired videos. Second, we reveal the reasons behind VLMs' poor performance on physical implausibility based on comprehensive experiments. Third, we propose a lightweight contrastive fine-tuning strategy, which enables VLMs to detect and explain physical implausibility in generated videos. Finally, we employ the fine-tuned PhyDetEx

to evaluate the physical plausibility of mainstream T2V models, providing new insights on their progress and challenges toward world modeling.

2 Related Work

Text-to-Video (T2V) Models. Recent large-scale T2V models have demonstrated significantly improved video fidelity, motion coherence, and instruction following capability [25, 26, 28, 38, 40, 48, 58, 60], making them a promising path toward building world models [11, 41, 57] that can capture the causal dynamics of the real world. However, despite the impressive improvement in video quality, these models often violate basic physical laws [24,55,66,68], such as object penetration, momentum inconsistency, or energy non-conservation. Such implausible generations even under semantically normal prompts are typically unacceptable for practical applications, which necessites the need to evaluate the physical plausibility beyond visual quality.

Physical Implausibility Detection. Given the prevalence of physically implausible content in T2V outputs, several approaches have been proposed for detecting it. One line of work trains end-to-end detectors with annotated datasets [5, 6, 16, 20, 53], assigning a physical plausibility score to generated videos. Another direction defines a set of rule-based prompts with known physical attributes and checks whether the generated content satisfies these predefined conditions [7,36,55,59,67]. Furthermore, Morpheus [63] proposes a physics-informed benchmark that evaluates generation models via real-world physical experiments and conservation-law metrics, offering quantitative and interpretable assessment of physical plausibility. However, both paradigms face key limitations: end-to-end detectors often lack interpretability, while rule-based methods are restricted to predefined prompt scenarios. Hence, a more generalizable and interpretable method is required to measure models' adherence to physical laws. In contrast, Vision-Language Models [3, 9, 10, 33], which are trained on massive multimodal data [13, 29] and are capable of producing textual explanations [19, 34, 43, 44], offer a promising alternative for interpretable and generalizable detection.

Vision-Language Models (VLMs). VLMs have achieved remarkable progress in recent years [3, 9, 27, 30, 33], excelling not only in traditional vision tasks such as visual question answering [18, 22, 43] and captioning [31, 47] but also in complex reasoning tasks like visual grounding [**?**, 49, 51] and generation [15, 62, 65]. Nonetheless, prior studies reveal that even strong VLMs struggle to detect physically implausible events in generated content [4, 54]. Impossible Videos [4] is the first work to evaluate VLMs' understanding of such implausibility, establishing a benchmark to assess their detection ability. However, it focuses solely on videos generated by implausible prompts and cannot reflect real-world T2V failures caused by normal prompts. These limitations motivate us to explore how to equip VLMs with the capability to not only detect physically implausible events and explain the violated physical principles, but also identify real-world T2V failures, thereby benchmarking the physical plausibility of current T2V models.

3 Preliminary Investigation

As introduced in Sec. 1, prior studies [4, 54] have shown that even powerful VLMs struggle to detect physically impossible or implausible content generated by T2V models. However, it is unclear whether these failures stem from a lack of understanding of physical plausibility, or are caused by the biases caused by training predominantly on physically plausible real-world data. To find out the reason and provide guidance for future improvements, we conduct a series of preliminary analyses. Specifically, we evaluate VLMs under three different prompt configurations with progressively increasing prior information based on the Impossible Videos benchmark [4]. Since all implausible videos in Impossible Videos are generated, we define three conditions: **(C1)** the model is not informed that the video is generated (our primary evaluation setting); **(C2)** the VLMs are told that the video may be generated; and **(C3)** the VLMs are explicitly told that the video is generated by a T2V model.

As shown in Fig. 2 (a), the detection performance varies substantially across these different conditions. The accuracy of identifying physically implausible content consistently increases from **C1** to **C3**, whereas the accuracy for physically plausible videos decreases accordingly. Furthermore, VLMs not only improve in detection accuracy but also produce **more precise explanations** for the implausible events. These findings reveal that VLMs inherently possess a latent understanding of physical laws, but their reasoning is heavily influenced by the assumed authenticity of the type of input video. As shown in Fig. 2 (b), when VLMs presume the video is real (as most of their training data are), they tend to overlook violations of physical plausibility. Conversely, when explicitly informed that the video is generated, they adopt a more critical stance and successfully identify physical implausibility. We provide more preliminary experiments with different VLMs and details in Sec 5 and **Supplemental Material A**.

4 Methodology

4.1 Construction of PID

Building upon our preliminary findings, we observe that VLMs indeed possess the capability to understand physically implausible content in generated videos. However, because their training data predominantly consist of physically plausible, real-world content, they tend to assume that any given video is real, and thus free of anomalies. Conversely, they often over-associate generated videos with implausibility. To mitigate this bias, we build the **P**hysical **I**mplausibility **D**etection (PID) dataset to help VLMs detect implausible content within generated data rather than merely distinguishing between real and generated videos.

Training Split. As shown in Fig. 3(a), the training split of PID is built from the real-world T2V dataset, VIDGEN-1M [45]. The goal of PID's training split is twofold: (1) it should not require a massive scale, since VLMs already exhibit a latent understanding of physical plausibility; and (2) it should explicitly expose

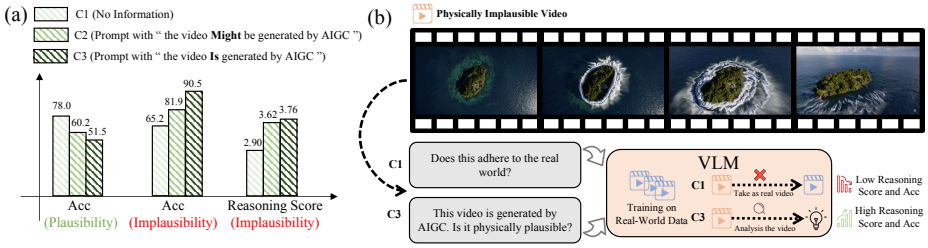

Fig. 2: (a) Results of preliminary experiments in the ImpossibleVideos under three prompting conditions (**C1**–**C3**) based on InternVL2.5 26B. As the prompt provides progressively stronger hints that the video may be generated by an AIGC model, the VLM achieve notably higher accuracy in detecting *physically implausible* videos and generate more accurate reasoning (higher reasoning scores). However, their accuracy on *physically plausible* videos decreases accordingly. These trends indicate that VLMs possess an implicit understanding of physical plausibility, yet their judgments are strongly biased by the type of input video they consider (real or generated). (b) Since most of its training data comes from the real world, VLM tends to assume that the input is real-world videos without providing any information.

the models to physical implausibility patterns that are absent from their pretraining data, but the overall content of the data should still be highly relevant to the real world, thereby breaking the shortcut that equates "generated videos" with "physically implausibile videos".

To construct the dataset, we first select videos from VIDGEN-1M that involve clear physical interactions with their corresponding captions. We then employ an LLM to rewrite each caption such that it contains a physically implausible description while preserving all other contextual elements. A T2V model is then used to generate videos based on these rewritten captions. This process produces video pairs where the physically plausible videos originate from VIDGEN-1M, and the corresponding physically implausible videos are generated from the modified caption. Based on our preliminary investigation, VLMs can reliably identify implausible content when provided with prior information. We then instruct the VLM that the videos are generated by a T2V model and use it to filter out pairs where the physically implausible content is accurately reflected in the generated video. Through this process, we obtain 2,588 paired videos.

This construction pipeline offers two major advantages. First, by generating physically implausible videos from the real-world ones, the PID training split closely mirrors the true distribution of physical content and avoids biases introduced by hand-defined "implausibility classes". Second, as illustrated in Fig. 3(a), during caption rewriting, we modify only the physical aspect of the scene while keeping all other elements unchanged. Consequently, each video pair differs only in physical plausibility, forming a contrastive structure that isolates the target concept. This ensures that the generated videos remain semantically similar to their real counterparts, reducing the domain gap from the VLMs' training data

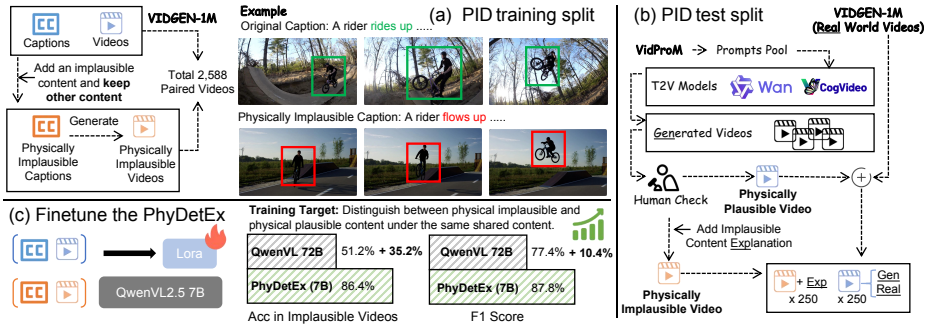

Fig. 3: Overview of the construction pipeline of the PID dataset and the training process of PhyDetEx. (a) The PID training split. The training split includes 2,588 paired videos, where each implausible video is generated by rewriting the caption of a real-world video to describe an implausible event while keeping other content unchanged. (b) The PID test split. The test set consists of physically implausible and physically plausible videos. The implausible subset is collected from videos generated by multiple T2V models based on physically plausible prompts, where human annotators identify implausible events and provide textual explanations. The plausible subset combines generated and real-world videos verified to contain no physical violations, thereby eliminating the shortcut of distinguishing between generated and real videos. (c) Training the PhyDetEx. Using the PID training split, we finetune the base VLM via LoRA adaptation to distinguish between physically plausible and implausible events within the same video contexts. The resulting model, PhyDetEx, achieves substantial improvements in detecting physically implausible content.

and eliminating the shortcut of detecting the physical implausibility. The advantages of such a design are further analysed in **Supplemental Material B**.

Test Split. To verify whether the model truly acquires the ability to detect physical implausibility, we introduce the test split of our PID. While Impossible Videos [4] offers an initial benchmark to evaluate the physical implausibility detection ability of VLMs, it suffers from two key limitations. First, most of its videos are derived from explicitly physically implausible prompts, which diverge from realistic scenarios, where users typically input plausible prompts, but T2V models still produce physically implausible videos. Second, the physically plausible videos of Impossible Videos consist entirely of real-world videos, creating a shortcut: VLMs may simply distinguish between generated and real videos instead of genuinely reasoning about physical plausibility.

To address these issues, we construct the PID test split, specifically designed to remove such shortcuts and more faithfully assess VLMs' understanding of physical laws. As illustrated in Fig. 3(b), the construction of test split involves several stages. We first sample prompts from VidProM [52], a large-scale prompts dataset crawled from the real user community, filtering out those containing inherently unrealistic or science fictional elements (*e.g.*, cartoon, futuristic, magical), as well as prompts lacking physical interactions (*e.g.*, static scenes). Using these filtered prompts, we generate videos via multiple T2V models and manually identify those exhibiting physical implausibility, accompanied by textual

explanations describing the physically implausible content. For the physically plausible subset, to eliminate the "generated vs. real" shortcut, we deliberately avoid relying solely on real-world videos. Instead, we combine physically plausible generated videos with real-world videos from VIDGEN-1M [45], filtered to be under 10 seconds and contain clear physical interactions. Following this pipeline, the PID test split comprises 250 physically plausible and 250 physically implausible videos, resulting in a total of 500 balanced test samples.

We provide detailed statistics and more details of the construction pipeline of our proposed PID train split and test split in **Supplemental Material C**.

4.2 Tuning the VLM

With the constructed PID dataset, our objective is to tune the VLM so that it can accurately detect *physically implausible* content in the generated videos. As discussed in Sec. 3, VLMs tend to treat "real" as "physically plausible" and "generated" as "physically implausible". Our tuning should remove this shortcut and enable the VLMs to reason about whether a video contains physically implausible content.

Notation. Denote by v_r a real-world video and t_r its corresponding caption with physically plausible events. We first rewrite t_r using an LLM, producing a modified caption t_g where the physically plausible event is changed to a physically implausible event, while keeping all other contextual elements the same:

$$t_g = \mathcal{R}_{\text{phys}}(t_r), \tag{1}$$

where $\mathcal{R}_{\text{phys}}(\cdot)$ denotes the rewriting operation. Then, a physically implausible video v_g is generated from the rewritten caption t_g using a T2V model $\mathcal{G}$:

$$v_g = \mathcal{G}(t_g). \tag{2}$$

The content in videos v_r and v_g can be partitioned into two parts, one part about the physically plausible/implausible event, and another shared part with the same content (*e.g.*, environment, background objects, scene layout), denoted by c_{share}. Then, we can express the content of the two videos as follows:

$$v_r = (c^r_{\text{phys}}, c_{\text{share}}), \quad v_g = (c^g_{\text{phys}}, c_{\text{share}}), \tag{3}$$

where c^r_{phys} and c^g_{phys} represent the content of the physically plausible and implausible events in the real and generated videos, respectively. The resulting paired samples can then be represented as:

$$(v_r, t_r), \quad (v_g, t_g), \tag{4}$$

where (v_r, t_r) represents a physically plausible video from the real world, and (v_g, t_g) represents the corresponding physically implausible video generated from the modified caption while preserving the other content.

Training Objective. Given paired videos (v_r, v_g), the goal of our training is to teach the VLM to correctly distinguish the content of each video as physically plausible or implausible. Let $f_\theta(\cdot)$ denote the prediction of a VLM model on the physical plausibility of a video:

$$y_r = f_\theta(v_r), \quad y_g = f_\theta(v_g), \tag{5}$$

where $y_r, y_g \in \{\text{physically plausible}, \text{physically implausible}\}$. The model is trained with the target such that, for each paired sample videos, the predicted labels should correctly reflect the plausibility of the physical content:

$$\begin{cases} y_r = \text{physically plausible}, \\ y_g = \text{physically implausible}. \end{cases} \tag{6}$$

In other words, the VLM is encouraged to rely solely on the content c_{phys} to make its decision, while ignoring the shared content c_{share}. This ensures that the model does not exploit shortcuts based on background, environment, or other unrelated content of the underlying physical event. There are two key motivations for this design. First, physical event is a highly diverse and context-dependent phenomenon that is difficult to categorize explicitly. By generating physically implausible samples through caption modifications of real-world videos, the resulting content remains within the distribution of real-world motions, preserving meaningful context dependencies; for instance, surfing motions only make sense in scenes containing water. Second, as discussed in Sec. 3, our preliminary investigation indicates that VLMs may develop a shortcut correlating generated videos with physical implausibility and real-world videos with physical plausibility, rather than reasoning about the physics. Our design minimizes cues from generation artifacts and encourages the model to focus on the physical event itself, thereby mitigating such shortcuts.

Language Modeling Supervision. Because the physically implausible videos v_g are generated from rewritten captions t_g with known physical implausibility, the model can be directly supervised to predict both the correct plausibility label and its corresponding reasoning. Similar to the current training objective of VLMs and LLMs, the entire training can be unified under a single language modelling (LM) objective. For each video–caption pair (v, t), the VLM is required to generate a textual sequence $\mathbf{s} = [s_1, s_2, \ldots, s_T]$, where the first token s_1 represents the physical plausibility judgment (`[YES]` or `[NO]`), and the remaining tokens $\{s_2, \ldots, s_T\}$ form a natural-language explanation supporting this judgment. Given the visual and textual inputs, the model is trained autoregressively using the standard negative log-likelihood loss:

$$\mathcal{L}_{\text{LM}}(\mathbf{s} \mid v, t) = -\textstyle\sum_{u=1}^{T} \log p_\theta(s_u \mid s_{<u}, v, t), \tag{7}$$

where θ denotes the trainable parameters of the VLM.

This unified LM-based supervision activates the VLM's latent capability to detect physical implausibility and leverages its inherent strength in generating textual explanations, enabling the model to serve as both a reliable detector and an interpretable reasoner for physical plausibility.

Model	Impossible Video				PID Test			
	Acc. Impl.	**Acc. Plaus.**	**F1 Score**	**Reasoning Score**	**Acc. Impl.**	**Acc. Plaus.**	**F1 Score**	**Reasoning Score**
Random Guess	50.0	50.0	47.4	–	50.0	50.0	50.0	–
InternVL2.5 4B [8]	61.3	69.3	69.0	2.72	61.6	84.4	75.8	2.26
InternVL2.5 8B [8]	52.3	76.9	71.2	2.32	53.2	90.8	76.4	2.13
InternVL2.5 26B [8]	65.2	78.0	73.2	2.89	58.4	92.8	79.2	2.16
LLaVA One-Vision [27]	29.3	81.5	68.0	1.70	36.8	94.0	73.1	1.75
MiniCPM-V 4.5 8B [61]	21.3	**98.9**	75.0	1.91	10.4	**97.6**	68.0	1.52
Qwen2.5VL 3B [3]	38.7	84.3	71.8	2.36	45.2	88.0	72.5	2.30
Qwen2.5VL 7B [3]	64.9	69.8	70.3	1.73	57.2	86.4	75.4	1.66
Qwen2.5VL 32B [3]	63.2	84.5	78.7	1.84	35.6	96.4	73.9	1.70
Qwen2.5VL 72B [3]	51.2	88.5	77.4	3.22	54.0	91.2	76.9	3.13
Qwen3VL 8B [2]	47.3	97.8	81.1	3.20	51.6	93.6	77.4	3.13
Qwen3VL 30B A3 [2]	63.2	84.4	78.7	3.71	58.4	90.0	77.7	3.28
PhyDetEx (ours, 7B)	**86.4**	87.1	**87.8**	**4.20**	**75.6**	89.2	**83.5**	**4.24**

Table 1: Quantitative comparison of physical implausibility detection results across different VLMs on the Impossible Videos and our PID test split. The **bold** indicates the best performance. We report four metrics for each model: accuracy on implausible videos (**Acc. Impl.**), accuracy on plausible videos (**Acc. Plaus.**), overall detection performance (**F1 Score**), and explanation quality (**Reasoning Score**). Our proposed PhyDetEx achieves the best performance on both datasets, significantly surpassing existing VLMs in F1 Score and Reasoning Score, demonstrating its superior capability in detecting and reasoning about the physical implausibility in generated videos.

5 Experiments

5.1 Experiments Setup

Training Details. We adopt QwenVL2.5-7B as the base model and fine-tune it using LoRA on our PID training split for three epochs. The learning rate is set to 1e-4 with a cosine learning rate scheduler. We set the LoRA rank to 8 and apply it to all target modules to ensure comprehensive adaptation. To further validate the effectiveness of our PID train split and training strategy in enhancing the model's understanding of physical plausibility, we conduct additional analyses and experiments in **Supplemental Material B**.

Evaluation Datasets. We evaluate our proposed PhyDetEx on two datasets: Impossible Videos [4] and our PID test split. For Impossible Videos, to better align with our focus on physical plausibility detection, we select the videos under the **Physical Laws** category, comprising 535 physically implausible videos. For the physically plausible subset, we adopt all real-world videos, resulting in 650 physically plausible samples. Our PID test split contains 250 physically implausible and 250 physically plausible videos. More details of our PID dataset are described in **Supplemental Material C**.

Model	C1			C2			C3		
	Acc. Impl	Acc. Plaus	Reasoning Score	Acc. Impl	Acc. Plaus	Reasoning Score	Acc. Impl	Acc. Plaus	Reasoning Score
InternVL2.5 8B	52.3	76.9	2.32	69.35	74.77	2.81	87.29	48.31	3.11
InternVL2.5 26B	65.23	78	2.9	81.87	60.15	3.62	90.47	51.54	3.76
Qwen2.5VL 7B	53.08	73.85	1.73	62.43	71.85	1.88	64.86	69.85	1.69
Qwen2.5VL 32B	45.98	94.62	1.84	52.52	86.92	2.04	60.75	82.46	2.06
Qwen2.5VL 72B	51.21	88.46	3.22	49.72	89.08	3.32	56.07	87.08	3.38

Table 2: Preliminary investigation results across different VLMs under conditions C1–C3. For each model, we report *Acc. Impl*, *Acc. Plaus*, and *Reasoning Score*. All models exhibit substantial performance shifts as the prompt condition changes.

Evaluation Metrics. We compare our PhyDetEx with several state-of-the-art VLMs using four metrics, specifically: (1) **Accuracy Implausible (Acc. Impl.)**: accuracy on physically implausible videos; (2) **Accuracy Plausible (Acc. Plaus.)**: accuracy on physically plausible videos; (3) the **F1 Score**: the overall detection performance balancing between the two accuracies; and (4) the **Reasoning Score**: the correctness and depth of the model's explanation for the detected physical implausibility.

The inclusion of the F1 Score is crucial, as many VLMs tend to exploit shortcuts, achieving high accuracy on the physically plausible class while failing on the other. For instance, MiniCPM-V 4.5 attains an exceptionally high Accuracy Plausible of 98.9% on ImpossibleVideos, but this coincides with a very low Acc Implausible, indicating a strong bias toward assuming all videos are physically plausible. Thus, the F1 Score serves as a more reliable indicator of true detection capability. The Reasoning Score further evaluates whether the model genuinely understands the underlying physical inconsistency rather than relying on superficial correlations. Detailed metric definitions and evaluation protocols are provided in the **Supplemental Material D**.

5.2 More Results of Preliminary Investigation

In addition to InternVL2.5-26B (see Fig. 2 (a)), we also evaluate several other VLMs, as shown in Tab. 2. The results show that across nearly all VLMs, the output changes substantially when provided with different prompts under conditions C1–C3: *Acc. Impl* and *Reasoning Score* increase significantly, while *Acc. Plaus* drops significantly. This suggests that VLMs tend to assume that a generated video is physically plausible, rather than leveraging their broad pretraining knowledge to assess the physical consistency of the video content. These observations further motivate the design and construction of our PID dataset.

5.3 Experimental Results

As shown in Tab. 1, on both Impossible Videos and our PID test split, most VLMs exhibit a notable bias when detecting physically implausible videos. For example, MiniCPM-V 4.5 achieves 97.6% Acc. Plaus. but only 10.4% Acc. Impl. on the PID test split. Similarly, LLaVA-One-Vision shows a comparable bias,

leading to high Acc. Impl. but low F1 Scores. It is worth noting that, the recent QwenVL3 8B and QwenVL3 30B-A3 models, although having smaller model sizes than QwenVL2.5 32B and 72B, achieve the highest F1 Scores among all baselines on both our PID test and the Impossible Video benchmark. This observation suggests that current VLM models' capacity is sufficient for physical implausibility detection, but the lack of implausible content in pretraining data prevents them from fully unlocking their potential in detecting physical implausibility. With architectural and training improvements, recent VLMs can demonstrate strong physical reasoning ability even at smaller scales.

Nevertheless, our tuned PhyDetEx substantially outperforms all competitors across Acc. Impl., F1 Score, and Reasoning Score on both datasets, achieving state-of-the-art results. PhyDetEx does not reach the highest Acc. Plaus., which is largely due to current VLMs' strong bias toward predicting plausibility. Therefore, the F1 Score is the most appropriate and balanced measure for this task. Moreover, the superior Reasoning Score confirms that PhyDetEx can not only detect physical implausibility but also provide accurate, interpretable explanations leveraging the VLM's inherent reasoning capabilities. We also provide performance analysis of the baseline in **Supplemental Material E**.

5.4 Case Study

In Fig. 5, we present two representative examples from our PID test split, comparing the outputs of PhyDetEx with several recent VLMs. The first case depicts a dolphin floating above the sea surface. LLaVA-OneVision and MiniCPM-V misinterpret this as a dolphin jumping out of the water, thus considering it physically plausible. They fail to recognize that the dolphin's height remains nearly constant over time, showing no effect of gravity. In contrast, PhyDetEx correctly identifies the scene as physically implausible and explicitly reasons that the object does not fall under gravity.

The second case is more complex: a woman jumps from a jungle cliff toward a river's surface but remains suspended above it without descending. Similar to the first example, MiniCPM-V again overlooks the implausibility and interprets the scene as a normal jumping motion, while QwenVL2.5-72B focuses on general scene description rather than the physical implausibility. In comparison, PhyDetEx successfully identifies the implausible subject and provides a concise, accurate explanation for its abnormality. We also provide additional qualitative examples and a detailed discussion of the failure cases with the corresponding reasoning in **Supplemental Material F**.

Beyond case study, we further use attention maps to examine whether fine-tuning on the PID training split mitigates shortcut behavior. As shown in Fig. 4, the attention maps provide a direct diagnostic signal: after fine-tuning, PhyDetEx focuses more on the abnormal regions, such as water and bubbles, than the baseline on the same example.

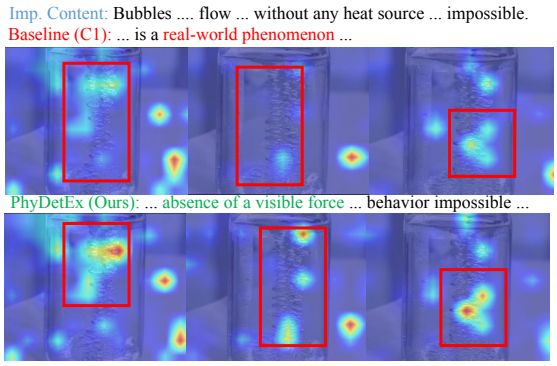

Fig. 4: Baseline vs. PhyDetEx attention. PhyDetEx highlights abnormal regions (water, bubbles); baseline is more dispersed.

T2V Model	Phy plausible Evaluation	
	Rate	*Score*
Open-Source		
CogVideoX1.5	52.0	5.35
HunyuanVideo	60.0	8.73
WanX2.1 1.3B	66.0	16.54
WanX2.1 14B	62.0	10.84
WanX2.2 14B	66.0	12.96
Clsoed-Source		
Veo3.1	72.0	12.98
Sora2.0	86.0	32.62

Table 3: Physical plausibility evaluation results of videos generated by various T2V models, benchmarked with our PhyDetEx. The closed-source model Sora2.0 shows the best performance, significantly higher than the second best player, Veo3.1.

5.5 Benchmarking Modern T2V Models

Having validated that our PhyDetEx can effectively detect the physically implausibility in generated videos, we leverage it to assess the ability of recent state-of-the-art T2V models in generating physically plausible videos. Specifically, we first extract 50 prompts involving clear physical interactions from VIDGEN-1M [45]. These prompts are then used to generate videos with a variety of T2V models, including open-source models—*CogVideoX1.5 5B* [60], *HunyuanVideo* [26], *WanX2.1 5B*, *WanX2.1 14B*, and *WanX2.2 14B* [48]—as well as the closed-source systems—*Veo3.1* [17] and *Sora2.0* [39]. Finally, we apply PhyDetEx to detect whether the generated videos exhibit physical implausibility.

PhyDetEx produces not only binary judgments but also token-level logits. Based on these outputs, we compute two complementary indicators: (1) the **Plausibility Rate**, defined as the proportion of videos judged to be physically plausible; and (2) a fine-grained **Plausibility Score** derived from the logits corresponding to the plausibility prediction, providing a continuous measure

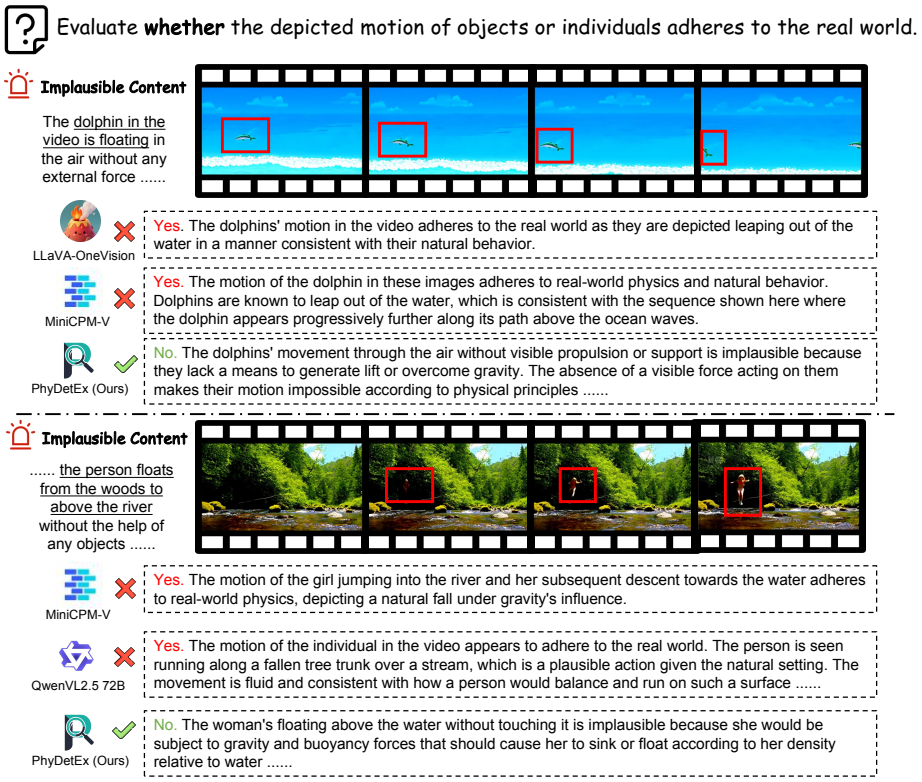

Fig. 5: Qualitative comparison between our PhyDetEx and recent VLMs on detecting physical implausibility. We illustrate two representative cases from our PID test split. In the first case (top), the dolphin remains floating above the sea surface without descending, violating gravity. LLaVA-OneVision and MiniCPM-V misinterpret the scene as a normal leap. By contrast, our PhyDetEx correctly identifies the physical implausibility and attributes it to the lack of gravitational influence. In the second case (bottom), a woman jumps from the woods toward the river but remains suspended midair. Other VLMs again regard this as plausible or irrelevant to physics, whereas our PhyDetEx identifies the implausible subject and provides the correct reasoning that her motion defies gravity and buoyancy.

of physical realism. Detailed information on the benchmarking procedure for T2V models is provided in the **Supplemental Material G**. In particular, for videos judged as "Physically Plausible" by PhyDetEx, the corresponding logits are added to the score, whereas for videos judged as "Physical Implausible", the logits are subtracted, resulting in an overall cumulative score.

The results are shown in Tab. 3. We see that the closed-source T2V system Sora2.0 achieves significantly higher *Rate* and *Score* than other models, demonstrating its superior understanding of physical laws. Veo3.1 is the second best but lags much behind Sora2.0. Among open-source models, *WanX2.1 1.3B* attains

the same *Rate* as *WanX2.2 14B* and achieves the highest *Score*. Our analysis indicates that although *WanX2.1 1.3B* produces lower visual quality compared to larger models, this does not reflect a deficiency in its physical reasoning capability. Rather, its tendency to generate simpler actions naturally reduces the likelihood of producing physically implausible content.

Besides benchmarking existing T2V models, we also leverage PhyDetEx to generate positive and negative samples for these models, enabling physical-awareness Direct Preference Optimization (DPO [42]) to improve them. Detailed implementation and results are provided in **Supplemental Material H**.

6 Conclusion

In this work, we first investigated why current VLMs struggled to identify physical implausibility in generated videos. We then proposed to construct PID, a dataset specifically designed for the physical implausibility detection task, consisting of a train and a test splits. Based on the train split, we developed PhyDetEx, a VLM-based detector capable of not only identifying but also explaining physical implausibility. PhyDetEx achieved SOTA performance on both our test split and the public Impossible Video benchmark. Furthermore, leveraging PhyDetEx, we evaluated a range of modern T2V models in terms of their ability to generate physically plausible videos. Our findings revealed that the powerful closed-source models have made significant progress in producing physically consistent videos, but the open-source models still lagged substantially behind.
Limitations. Although our trained PhyDetEx model performs well in detecting physical implausibility in generated videos, the diversity of physical events is vast. Therefore, PhyDetEx still exhibits hallucinations in non-visual or complex physical events. We provide several failure cases and discussion of PhyDetEx in **Supplemental Material F**. We believe that adding more modality information from the video, such as audio and additional information about objects in the video, rather than relying solely on purely visual modalities, can better help VLMs understand and detect physical implausibility.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006 and 23hytd006-2, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University, and in part by The Major Key Project of PCL (Grant No. PCL2025A17).

References

1. An, R., Yang, S., Guo, Z., Dai, W., Shen, Z., Li, H., Zhang, R., Wei, X., Li, G., Wu, W., et al.: Genius: Generative fluid intelligence evaluation suite. arXiv preprint arXiv:2602.11144 (2026)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bai, Z., Ci, H., Shou, M.Z.: Impossible videos (2025), https://arxiv.org/abs/2503.14378
5. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)
6. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025)
7. Chen, Y., Guo, X., Shi, Z., Song, Z., Zhang, J.: T2vworldbench: A benchmark for evaluating world knowledge in text-to-video generation (2025), https://arxiv.org/abs/2507.18107
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
9. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
11. Cho, J., Puspitasari, F.D., Zheng, S., Zheng, J., Lee, L.H., Kim, T.H., Hong, C.S., Zhang, C.: Sora as an agi world model? a complete survey on text-to-video generation. arXiv preprint arXiv:2403.05131 (2024)
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al., L.M.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
13. Cui, E., He, Y., Ma, Z., Chen, Z., Tian, H., Wang, W., Li, K., Wang, Y., Wang, W., Zhu, X., Lu, L., Lu, T., Wang, Y., Wang, L., Qiao, Y., Dai, J.: Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o (2024), https://sharegpt4o.github.io/
14. Cui, S., Guo, J., An, X., Deng, J., Zhao, Y., Wei, X., Feng, Z.: Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 950–959 (June 2024)

15. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
16. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. arXiv preprint arXiv:2504.00983 (2025)
17. Google DeepMind: Veo - google deepmind (2025), https://deepmind.google/models/veo/ [2025.09.08]
18. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
19. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14953–14962 (2023)
20. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., Wang, K., Do, Q.D., Ni, Y., Lyu, B., Narsupalli, Y., Fan, R., Lyu, Z., Lin, B.Y., Chen, W.: VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 2105–2123. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.127, https://aclanthology.org/2024.emnlp-main.127/
21. Huang, S., Wu, J., Zhou, Q., Miao, S., Long, M.: Vid2world: Crafting video diffusion models to interactive world models (2025), https://arxiv.org/abs/2505.14357
22. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
23. Jiang, C., Heng, Y., Ye, W., Yang, H., Xu, H., Yan, M., Zhang, J., Huang, F., Zhang, S.: Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. arXiv preprint arXiv:2505.16192 (2025)
24. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Gao, H., Jiashi, F.: How far is video generation from world model? – a physical law perspective. arXiv preprint arXiv:2406.16860 (2024)
25. Kling: Klingai: Image to video. https://app.klingai.com/global (2025), accessed: 2025-09-08
26. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
27. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
28. Li, X., Zhou, D., Zhang, C., Wei, S., Hou, Q., Cheng, M.M.: Sora generates videos with stunning geometrical consistency. arXiv preprint arXiv:2402.17403 (2024)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
30. Lin, W., Wei, X., An, R., Peng, G., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable mllms to compre-

hend what you want. In: International Conference on Learning Representations. vol. 2025, pp. 46374–46403 (2025)
31. Lin, W., Wei, X., An, R., Ren, T., Chen, T., Zhang, R., Guo, Z., Zhang, W., Zhang, L., Li, H.: Perceive anything: Recognize, explain, caption, and segment anything in images and videos. Advances in Neural Information Processing Systems **38**, 96231–96256 (2026)
32. Lin, W., Wei, X., Zhang, R., Zhuo, L., Zhao, S., Huang, S., Xie, J., Peng, G., Li, H.: Pixwizard: Versatile image-to-image visual assistant with open-language instructions. In: International Conference on Learning Representations. vol. 2025, pp. 38680–38711 (2025)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
34. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
35. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., Zhou, Y., Sun, D., Zhou, D., Zhou, J., Tan, K., An, K., Chen, M., Ji, W., Wu, Q., Sun, W., Han, X., Wei, Y., Ge, Z., Li, A., Wang, B., Huang, B., Wang, B., Li, B., Miao, C., Xu, C., Wu, C., Yu, C., Shi, D., Hu, D., Liu, E., Yu, G., Yang, G., Huang, G., Yan, G., Feng, H., Nie, H., Jia, H., Hu, H., Chen, H., Yan, H., Wang, H., Guo, H., Xiong, H., Xiong, H., Gong, J., Wu, J., Wu, J., Wu, J., Yang, J., Liu, J., Li, J., Zhang, J., Guo, J., Lin, J., Li, K., Liu, L., Xia, L., Zhao, L., Tan, L., Huang, L., Shi, L., Li, M., Li, M., Cheng, M., Wang, N., Chen, Q., He, Q., Liang, Q., Sun, Q., Sun, R., Wang, R., Pang, S., Yang, S., Liu, S., Liu, S., Gao, S., Cao, T., Wang, T., Ming, W., He, W., Zhao, X., Zhang, X., Zeng, X., Liu, X., Yang, X., Dai, Y., Yu, Y., Li, Y., Deng, Y., Wang, Y., Wang, Y., Lu, Y., Chen, Y., Luo, Y., Luo, Y., Yin, Y., Feng, Y., Yang, Y., Tang, Z., Zhang, Z., Yang, Z., Jiao, B., Chen, J., Li, J., Zhou, S., Zhang, X., Zhang, X., Zhu, Y., Shum, H.Y., Jiang, D.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model (2025), https://arxiv.org/abs/2502.10248
36. Meng, F., Liao, J., Tan, X., Lu, Q., Shao, W., Zhang, K., Cheng, Y., Li, D., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=dIjMswSzgF
37. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv preprint arXiv:2410.05363 (2024)
38. MiniMax: Hailuo ai: Transform idea to visual with ai (2025), https://hailuoai.video/ [2025.09.08]
39. OpenAI: Sora-2 (2025), https://openai.com/index/sora-2/
40. Pika Lab: Pika (2025), https://pika.art/ [2025.09.09]
41. Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., BAI, L., Zhang, R.: Worldsimbench: Towards video generation models as world simulators. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=j9pVnmulQm
42. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=HPuSIXJaa9
43. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. arXiv (2022)

44. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. arXiv preprint arXiv:2303.08128 (2023)
45. Tan, Z., Yang, X., Qin, L., Li, H.: Vidgen-1m: A large-scale dataset for textto-video generation. arXiv preprint arXiv:2408.02629 (2024)
46. Tong, C., Guo, Z., Zhang, R., Shan, W., Wei, X., Xing, Z., Li, H., Heng, P.A.: Delving into rl for image generation with cot: A study on dpo vs. grpo. Advances in Neural Information Processing Systems **38**, 115632–115655 (2026)
47. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
48. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
49. Wang, G., Wei, X., Liu, J., Zhang, R., Zhang, Y., Zhang, K., Chong, M., Zhang, S.: Mr-mllm: Mutual reinforcement of multimodal comprehension and vision perception. arXiv preprint arXiv:2406.15768 (2024)
50. Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al.: Wisa: World simulator assistant for physics-aware text-to-video generation. arXiv preprint arXiv:2503.08153 (2025)
51. Wang, S., Liu, S., Kuang, Y., Wei, X., Liu, Y., Li, Z., Man, Y., Chen, G., Tao, A., Liu, G., et al.: Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. arXiv preprint arXiv:2605.27365 (2026)
52. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. Thirty-eighth Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=pYNl76onJL
53. Wang, Y., Tan, Z., Wang, J., Yang, X., Jin, C., Li, H.: Lift: Leveraging human feedback for text-to-video model alignment. arXiv preprint arXiv:2412.04814 (2024)
54. Wang, Z., Ma, Q., Wan, W., Li, H., Wang, K., Tian, Y.: Is this generated person existed in real-world? fine-grained detecting and calibrating abnormal human-body. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21226–21237 (2025)
55. Wang, Z., Wei, X., Li, B., Guo, Z., Zhang, J., Wei, H., Wang, K., Zhang, L.: Videoverse: How far is your t2v generator from a world model? arXiv preprint arXiv:2510.08398 (2025)
56. Wei, X., Cen, K., Wei, H., Guo, Z., Li, B., Wang, Z., Zhang, J., Zhang, L.: Mico-150k: A comprehensive dataset advancing multi-image composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29695–29706 (2026)
57. Xiang, J., Liu, G., Gu, Y., Gao, Q., Ning, Y., Zha, Y., Feng, Z., Tao, T., Hao, S., Shi, Y., Liu, Z., Xing, E.P., Hu, Z.: Pandora: Towards general world model with natural language actions and video states (2024)
58. Xiao, J., Yang, C., Zhang, L., Cai, S., Zhao, Y., Guo, Y., Wetzstein, G., Agrawala, M., Yuille, A., Jiang, L.: Captain cinema: Towards short movie generation. arXiv preprint arXiv:2507.18634 (2025)

59. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18826–18836 (2025)
60. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
61. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
62. Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., Yan, H., Fu, J., Gui, T., Sun, T., Jiang, Y.G., Qiu, X.: AnyGPT: Unified multimodal LLM with discrete sequence modeling. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9637–9662. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.521, https://aclanthology.org/2024.acl-long.521/
63. Zhang, C., Cherniavskii, D., Tragoudaras, A., Vozikis, A., Nijdam, T., Prinzhorn, D.W., Bodracska, M., Sebe, N., Zadaianchuk, A., Gavves, E.: Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. arXiv preprint arXiv:2504.02918 (2025)
64. Zhang, R., Wei, X., Jiang, D., Guo, Z., Zhang, Y., Tong, C., Liu, J., Zhou, A., Zhang, S., Peng, G., et al.: Mavis: Mathematical visual instruction tuning with an automatic data engine. In: International Conference on Learning Representations. vol. 2025, pp. 87955–87989 (2025)
65. Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Chen, C.L., Yan, S.: Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In: NeurIPS (2024)
66. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: Videorepa: Learning physics for video generation through relational alignment with foundation models. arXiv preprint arXiv:2505.23656 (2025)
67. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)
68. Zhu, Z., Wang, X., Zhao, W., Min, C., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., Zhang, C., et al.: Is sora a world simulator? a comprehensive survey on general world models and beyond. arXiv preprint arXiv:2405.03520 (2024)