VLIPP: Towards Physically Plausible Video Generation with Vision and Language Informed Physical Prior

Xindi Yang[1]*, Baolu Li[2]*, Yiming Zhang[2], Zhenfei Yin[4, 5]†, Lei Bai[3]†, Liqian Ma[6], Zhiyong Wang[5], Jianfei Cai[1], Tien-Tsin Wong[1], Huchuan Lu[2], Xu Jia[2]†

[1]Monash University [2]Dalian University of Technology [3]Shanghai Artificial Intelligence Laboratory
[4]Oxford University [5]The University of Sydney [6]ZMO AI

Abstract

Video diffusion models (VDMs) have advanced significantly in recent years, enabling the generation of highly realistic videos and drawing the attention of the community in their potential as world simulators. However, despite their capabilities, VDMs often fail to produce physically plausible videos due to an inherent lack of understanding of physics, resulting in incorrect dynamics and event sequences. To address this limitation, we propose a novel two-stage image-to-video generation framework that explicitly incorporates physics with vision and language informed physical prior. In the first stage, we employ a Vision Language Model (VLM) as a coarse-grained motion planner, integrating chain-of-thought and physics-aware reasoning to predict a rough motion trajectories/changes that approximate real-world physical dynamics while ensuring the inter-frame consistency. In the second stage, we use the predicted motion trajectories/changes to guide the video generation of a VDM. As the predicted motion trajectories/changes are rough, noise is added during inference to provide freedom to the VDM in generating motion with more fine details. Extensive experimental results demonstrate that our framework can produce physically plausible motion, and comparative evaluations highlight the notable superiority of our approach over existing methods. More video results and code are available on our Project Page: https://madaoer.github.io/projects/physically_plausible_video_generation/.

1. Introduction

Video diffusion models (VDMs) trained on large-scale video datasets have made remarkable progress in terms of realism, demonstrating significant potential for various content creation applications. Despite the absence of explicit geometric modeling, the generated videos still exhibit coherent spatial relationships among objects, rich textured details, and realistic lighting effects, including reflections and shadows. Such qualities often make the generated videos nearly indistinguishable from real-world footages. This drives the research community to explore the potential of VDMs as world models. However, they still struggle with understanding the physical laws of the real world and generating videos that adhere to these principles.

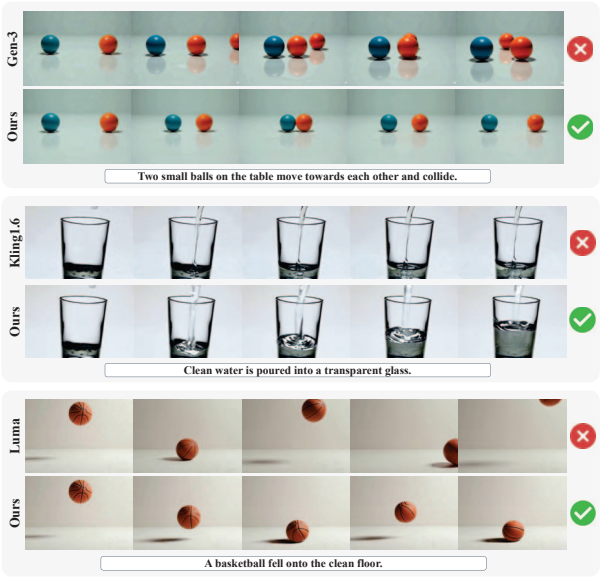

Figure 1. Existing commercial closed-source VDMs fail to generate physically plausible motion, whereas our video generation framework is able to achieve this by incorporating external physical prior knowledge.

Although existing VDMs can produce visually realistic videos, they fail to mimic the real-dynamic physical motions. As shown in Fig 1, even the current commercial closed-source VDMs struggle with the task of generating videos that conform to physical laws. PhyT2V [58] refines text prompts by incorporating detailed descriptions

* Equal contribution † Corresponding author

of physical processes to guide VDMs in generating physically plausible videos. However, despite being pretrained on internet-scale real-world video–text pairs, VDMs do not inherently understand physical laws. This limitation arises from the gap and ambiguity between text descriptions and the actual motion in the video[33]. Moreover, VDMs tend to overfit the training data rather than developing a general understanding of physical laws[21]. Inspired by the success of graphics-based physical rendering, some methods have guided VDMs to generate physically plausible videos using simulations from graphics engines[18, 29, 32, 56, 64]; however, these approaches rely on the physical effects that graphics engines can simulate and incur high computational costs.

The gap and ambiguity between text and real-world motion makes it difficult to enable physically plausible video generation through detailed text descriptions alone. Moreover, it is challenging to gather scalable physical data for training due to the abstractness and diversity of physical phenomena. Consequently, a viable approach could be to model abstract physical laws as conditions for diffusion models. However, it is less practical to explicitly model the physics equation for every kind of motion. Instead, we resort to current large foundation models for their ability to "understand" basic physics[6] and reason about physical phenomena based on the knowledge they extract. For example, given two colliding balls, the Large Language Model (LLM) can *approximately* predict the paths of the balls after collision. Inspired by this observation, we propose a novel video generation framework that employs a Vision Language Model (VLM) to predict the path/change during a physics event, described by a given image and a text prompt.

In this paper, we propose VLIPP, a two-stage approach to incorporate physics as conditions into VDM, enabling the generation of physically plausible motion. In the first stage, the VLM serves as a *coarse-level motion planner*, while a VDM serves as a *fine-level motion synthesizer*. The idea of stage one is to utilize the chain-of-thought and the physics-aware reasoning of VLM planning to ensure that coarse-level motion trajectories approximately follow real-world physics dynamics. In stage two, we can generate fine-level motion using an image-to-video diffusion model conditioned by the approximated path/change planned by VLM from stage one. Note that the approximated path/changes are not in the level to tell the speed or acceleration of the motion. We choose an existing image-to-video model [7] to accept our coarse-level path/change, by injecting noise to the motion path during both the training and inference phases. Notably, during the VLM planning stage, generating entire physically plausible motion trajectories is not required. Instead, we leverage the generative priors of VDM to produce fine-level physically plausible videos based on coarse-level motion trajectories provided by the VLM. So that the detail-level motion such as speed, acceleration, and vibration are left to the VDM to synthesize.

We evaluate our physically plausible video generation framework with two major video physics benchmarks and achieved satisfactory results. Furthermore, we discuss and analyze multiple insightful design choices in our video generation framework, such as employing a motion planner tailored for different physics categories, and enhancing the robustness of diffusion model to noisy trajectories. Our contributions are summarized as follows:

1. We introduce a novel image-to-video generation framework for generating physically plausible videos by leveraging the VLM and VDM priors, significantly outperforming the contemporary competitors.
2. We propose a novel chain-of-thought and physics-aware reasoning approach in VLM, along with random noise injection in the latent space during video generation, which effectively improves both the generation quality and physical plausibility.
3. We conduct a comprehensive experiments and user studies to demonstrate the effectiveness and generalization of our framework in physically plausible videos generation.

2. Related Work

2.1. Physically Plausible Visual Content Generation

Generating physically plausible videos offers substantial value to real-world applications such as scientific simulations[43], robotics [4, 59], and autonomous driving [10, 48]. Traditional graphics pipelines rely on simulation systems to model physical phenomena [36, 42]. Inspired by these approaches, recent studies [18, 29] have performed dynamic simulations in image space based on physical engines. Furthermore, some methods [56, 64] incorporate physical priors into 3D representations to enable the synthesis of physically plausible motions. However, these rule-based or solver-based simulators face limitations in expressiveness, efficiency, generalizability, and parameter tuning. Furthermore, these simulators require significant expertise, rendering them inaccessible and unfriendly for users.

In addition, Some studies have explored VDMs for generating physically plausible videos. Li *et al*. [24] models natural oscillations and swaying in frequency-domain. A downstream rendering module then animates static images based on the generated motion information. PhysDiff [62] introduces physical simulator as constrain into the diffusion process by projecting denoised motion of a diffusion step into a physically plausible motion. These methods mainly focus only on specific types of physical motion and do not establish a generalizable approach for generating physically plausible videos.

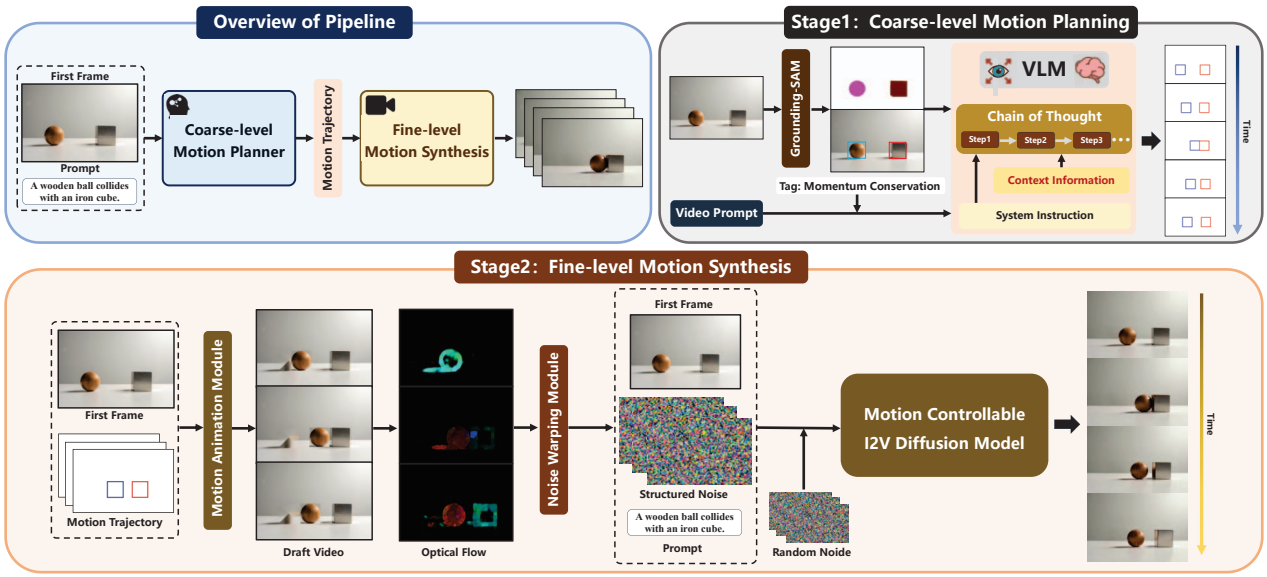

Figure 2. The illustration of our physically plausible image-to-video generation pipeline. Our pipeline consists of two stages. In the first stage, the VLM generates a coarse-grained, physically plausible motion trajectory based on the provided input conditions. In the second stage, We simulate a draft video using the predicted trajectory to provide the motion condition. We then extract the optical flow from this video and convert it into structured noise. These conditions are fed into a motion controllable image-to-video diffusion model, and ultimately generates a physically plausible video.

2.2. Motion Controllable Video Generation

Existing studies commonly provide one of the following three types of motion control: bounding box control [19, 23, 30, 46, 52], point trajectory control [34, 37, 40, 50, 54] and camera control [2, 15, 47, 61]. Bounding box control provides object motion guidance by generating a sequence of bounding boxes that track the object's position over time. Point-trajectory control offers motion cues through point-based trajectories, enabling drag-style manipulation. Camera motion control guides video generation using explicit 3D camera parameters, ensuring consistency and realistic viewpoint changes. However, these approaches prioritize motion control but often overlook physical plausibility. To address the limitation, we propose a novel framework for physically plausible video generation that incorporates physics as conditions into video diffusion models.

2.3. Generation based on VLMs Planning

VLMs have exhibited robust capabilities in visual understanding and planning [27, 39, 65]. Their strong performance in domains such as robot path planning and video understanding shows their ability in understanding the real physical world. Prior work has successfully leveraged LLMs to guide the layout of images or videos, yielding promising results [25, 53]. VideoDirectorGPT [26] leverages LLMs for fine-grained scene-by-scene planning, explicitly controlling spatial layout to generate temporally consistent long videos. Pandora [55] utilizes LLMs for real-time control through free-text action commands, achieving domain generality, video consistency, and controllability. However, these efforts have yet to address interactions with real-world physical phenomena, such as collision, fall, and melting.

Moreover, the absence of visual information can cause severe hallucination issues in language models for spatial planning tasks, leading to problems like overlapping object boundaries, disproportionate scaling, and incorrect planning [20, 57]. In this paper, we propose utilizing VLMs as coarse-level motion planners within the image space and incorporate physics-aware reasoning and Chain of Thought (CoT) [51] into the inference process.

3. Method

Task Fomulation. In this paper, our goal is to enable an image-to-video diffusion model to generate physically plausible videos. Since VDMs rely more on memory and case-based imitation and struggle to understand general physical rules[21], the key challenge is how to incorporate physical laws into the models. To achieve this, we need to identify a method to incorporate physical principles into the video diffusion framework. Given an image $I \in R^{H\times W\times C}$($H$ is height, W is width and C is the number of channels) and a text description d of possible events based on image I, our framework should infer a physics-compliant guidance as the input condition and synthesize a video that adheres to both physical laws and real-world dynamics.

Overall Pipeline. Overall pipeline of VLIPP is illustrated in Figure 2. In the first stage, the VLM conducts semantic

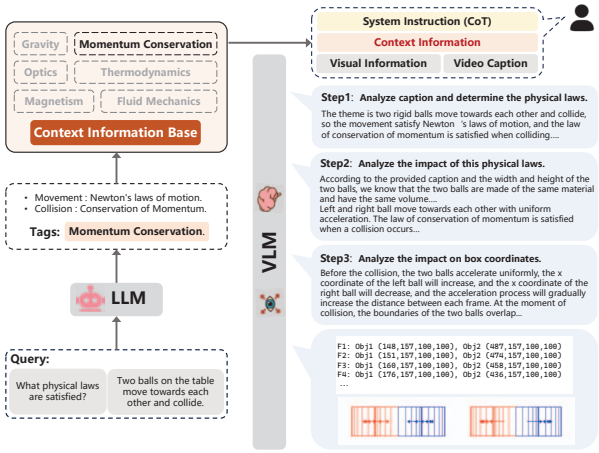

Figure 3. The illustration of chain-of-thought reasoning in the VLM for generating a coarse-grained motion trajectory. First off, the VLM determines the corresponding physical laws and its context for the given scene. Then, the VLM performs step-by-step reasoning to predict the physically plausible motions of objects in image space, leveraging physical context and chain-of-thought prompting. Finally, the VLM predicts bounding boxes according to real-world physics.

analysis and physical attribute analysis on the given image I and a description d to obtain the bounding boxes of the objects in the scene, denoted as $b_1, b_2, ..., b_n$, along with the applicable physical laws l. Next, the VLM infers possible future physical scenarios in the current scene to derive the coarse-level motion trajectories of $p_1, p_2, ..., p_t$ in the image space. Finally, we utilize an image-to-video diffusion model to synthesize the detailed dynamics in the video.

3.1. VLM as a Coarse-Level Motion Planner

Our motivation is to incorporate physical laws as constraints into a video diffusion model to enhance the physical plausibility of the generated videos. To achieve this, we must identify a method to inject physical laws into the video diffusion model. Given a video description and the first frame, the task at this stage is to generate coarse-level motion trajectory aligned with physical laws.

Scene Understanding. In the real world, most physical phenomena arise from interactions among objects and their motion trajectories. We first initiate the process by identifying and locating objects within a scene. Inspired by the recent studies [1, 9, 28] in VLMs for scene understanding, we employ GPT-4o [38] to recognize all objects that could be involved in physical phenomena as described in the text description d. These objects are subsequently detected and segmented using Grounded-SAM2 [41], yielding their bounding boxes. By leveraging the pretrained knowledge and common-sense reasoning capabilities of foundation models, we effectively determine the relevant objects in the scene.

Physical-Aware Recognition. To perform more effective reasoning in predicting the motion, it is necessary to determine what specific physical principle to apply in the given context. We utilize the pretrained prior of the LLM to determine the physical laws applicable to the current scene. Following the configuration in the physical benchmark [3, 31, 33], we currently classify common physical phenomena in videos into six categories: gravity, momentum conservation, optics, thermodynamics, magnetism, and fluid mechanics. Note that such list can be easily extended within our framework. Given a video description d, the LLM infers the physical law l that governs the current scene. We provide the specific physical context information for VLM to enhance its understanding of physical laws [11]. Detailed context design is presented in the Appendix.

Chain of Thought Reasoning in VLM. Given the physical law l, an image I and a video description d for the scene, we prompt the VLM to predict the future bounding box positions of objects within the image-space. We choose to predict in the image space for two primary reasons. Firstly, motion in image space aligns more with our subsequent video synthesizer. Secondly, image space dynamics can effectively represent a wide range of real-world motions [29].

At a given time t, the predicted position of i-th object bounding box b_t^i is denoted as $[x_t^i, y_t^i, w_t^i, h_t^i]$, where (x_t^i, y_t^i) represents its top-left coordinate; w_t^i & h_t^i denote its width and height, respectively. Governed by the physical law, the four values of the bounding box may change over time. The VLM reasons the bounding box positions of N future frames for every object o_i based on the condition. To help VLM better understand physical laws, we adapt a chain-of-thought [51] into its reasoning, to significantly enhance its reasoning capabilities. As shown in Figure 3, we formulate our analysis of physical phenomena in videos as step-by-step reasoning: beginning with broad conceptual ideas and progressing to a detailed and practical examination:

1. Given the physical law l and context information, the VLM analysis video caption and detail the physical law.
2. The VLM analyzes the potential interactions and movement of each object within the scene;
3. The VLM predicts the detailed changes in position and shape of the bounding box corresponding to each object over time.

Through the structured planning process, the VLM plans coarse-level motion trajectories for the objects, approximating real-world physics dynamics. In particular, our VLM infers the changes of object bounding boxes for next 12 frames, constrained by the token length limitation. To be compatible with the generation process of the chosen VDM in the next stage, these inferred 12 frames are further linearly interpolated to produce a total of 49 frames.

3.2. VDM Serves as a Fine-Level Motion Synthesizer

In the previous stage, the motion trajectory planned by the VLM is neither precise nor fully compliant with physical laws. On the other hand, while VDM may not be able to produce realistic global motion trajectories, it is able to generate sound motion in finer scale. In this stage, our key insight is that the VDM can refine the coarse-level motion to produce physically plausible motion that aligns with real-world dynamics with its powerful generative prior.

Motion Animation. To incorporate physical laws into the video diffusion model, we use the inferred coarse motion trajectory to guide the generation process of the diffusion model. Optical flow provides a unified representation of motion, and recent studies [7, 12, 13] have demonstrated its effectiveness in guiding diffusion models. Accordingly, we leverage the coarse-level motion trajectory to animate a draft motion sequence and derive the corresponding optical flow. Specifically, for each object o_i , we extract its bounding box from the first frame and move it to the bounding box location b_i specified by the draft motion trajectory. To animate the change of shape (e.g., due to compression or expansion), we resize object o_i according to the difference between o_i and o_{i+1} during inpainting. The draft motion video is generated as follows:

$$\hat{V}(t) = \text{Animation}(B, rs(o_0^0, b_t^0)...rs(o_0^i, b_t^i)) \quad (1)$$

where $\hat{V}(t)$ denotes the corresponding frame of inpainted video at timestep t, B denotes the inpainted background with the foreground object removed, o_i^0 denotes the i-th object at timestep 0, b_i^t represents the i-th bounding box at timestep t, and rs denotes resize function.

Structured Noise from Draft Video. Optical flow is an effective representation for guiding VDMs [12, 13]. Follow prior work [7, 8], we employ RAFT[44] to extract optical flow from the draft video and formulate it as structural noise, which retains Gaussian properties. Given the draft video $\hat{V}(t) \in \mathbb{R}^{F\times C\times H\times W}$, we calculate its per-frame optical flow to get a structured noise tensor $Q \in \mathbb{R}^{F\times C\times H\times W}$. The structured noise enables the VDM to generate videos that exhibit motion patterns closely aligned with those in the optical flow, thereby improving the realism of the output.

Noise Injection in Video Synthesis. We adopt Go-with-the-Flow [7] as our video synthesis model, a fine-tuned CogVideoX [60], which is designed to accept structured noise Q as input and synthesize videos that adhere to the implicit optical flow. The vanilla Go-with-the-Flow tends to tightly follow the provided structured noise Q. However, our Q is derived from a coarse-level motion trajectory and may not be sufficiently accurate to follow the physical laws of the real world. To address this limitation, we inject noise during the inference phase to give more flexibility to the VDM to generate detail-level motion changes as

$$Q_i = \frac{(1-\gamma)Q_i + \zeta\gamma}{\sqrt{(1-\gamma)^2+\gamma^2}} \quad (2)$$

where Q_i is structured noise at i-th frames, $\zeta \in \mathbb{R}^{C\times H\times W}$ is Gaussian noise and $\gamma \in [0, 1]$. We set $\gamma = 0.4$ for even frame index and $\gamma = 0.6$ for odd frame index.

With this approach, the VDM is able to generate motion deviate from the coarse-level motion trajectory whenever necessary for producing high-quality fine-level motion.

4. Empirical Analysis and Disscusion

In this section, we conduct extensive experiments to demonstrate the effectiveness of our video generation framework compared to existing methods. We evaluate our approach on two established benchmarks for physically plausible video generation. Our framework consistently achieves superior performance across all benchmarks.

4.1. Implementation Details

We propose a two-stage physically plausible image-to-video generation framework. In the first stage, we utilize ChatGPT-4o as the coarse-level motion planner. In the second stage, we utilize an open-source I2V model, Go-with-the-Flow [7], as a fine-level motion synthesizer. Unless otherwise specified, in all experiments, we generate each video with a resolution of 720×480 and 49 frames.

4.2. Benchmarks and Models

Traditional metrics in the visual domain, such as the Peak Signal-to-Noise Ratio (PSNR)[17], the Structural Similarity Index (SSIM)[49], the Learned Perceptual Image Patch Similarity (LPIPS)[63], the Fréchet Inception Distance (FID)[16] and the Fréchet Video Distance (FVD)[45], do not account for the physical realism of the generated videos [31, 33]. Recent studies have begun to address this limitation by developing benchmarks and metrics that evaluate physical realism. In this work, we adopt two benchmarks, described below.

PhyGenBench [31] categorizes physical properties into four domains: mechanics, optics, thermal, and material. It includes 27 physical phenomena, each governed by real world physical laws, reflected in 160 carefully designed text prompts. As PhyGenbench provides only text prompts, we adapt it to our image-to-video setting by generating a corresponding first frame for each prompt with FLUX[22]. We adhere to the predefined benchmark evaluation protocol, i.e., employing GPT-4o to assess the physical realism of the generated videos.

Physics-IQ [33] comprises 396 real-world videos spanning 66 distinct physical scenarios. For each scenario, videos

Figure 4. Visual comparisons of physically plausible video generation results from our framework, CogVideoX-I2V-5B [60], LTX-Video-I2V [14] and SVD-XT [5].

Model	Mechanics(↑)	Optics(↑)	Thermal(↑)	Material(↑)	Average(↑)
CogvideoX-T2V-5B	0.43	0.55	0.40	0.42	0.45
LTX-Video-T2V	0.35	0.45	0.36	0.38	0.39
OpenSora	0.43	0.50	0.44	0.37	0.44
PhyT2V	0.49	0.61	0.49	0.47	0.52
LLM-Grounding Video Diffusion	0.32	0.41	0.26	0.24	0.31
CogvideoX-I2V-5B	0.48	0.69	0.43	0.41	0.52
SVD-XT	0.46	0.68	0.48	0.41	0.52
LTX-Video-I2V	0.47	0.65	0.46	0.37	0.50
SG-I2V	0.52	0.69	0.51	0.39	0.54
Ours	**0.55**	**0.71**	**0.60**	**0.53**	**0.60**

Table 1. Quantitative results of VDMs on PhyGenBench.

are recorded from three different perspectives and filmed twice under identical conditions to eliminate randomness. This benchmark evaluates real-world physical phenomena, including collisions, object continuity, occlusion, object permanence, and fluid dynamics. This benchmark assesses physical realism from semantic and temporal perspectives, using semantic metrics and visual metrics to compare generated videos against the real-world reference videos.

Compared Models. In the context of text-to-video generation, we compare our framework with CogVideoX-T2V-5B [60], LTX-Video-T2V [14], and OpenSora [66]. Moreover, we evaluate our framework against PhyT2V [58], which enhances physical realism by iteratively refining the prompt. For the image-to-video generation scenario, our framework is evaluated alongside CogVideoX-I2V-5B [60], SVD-XT [5], and LTX-Video-I2V [14]. Additionally, we conducted experiments in the motion-controllable setting. In this setting, we leverage the motion trajectory predicted by the VLM as a condition to guide VDM generation. We benchmark our approach against image-to-video motion controllable model, SG-I2V [35] and text-to-video motion controllable model LLM-grounded Video Diffusion Models [25]. The experimental details are presented in the Appendix.

4.3. Quantitative Evaluation

We begin with an empirical study on PhyGenBench and Physics-IQ, comparing our framework against widely adopted open-source models in the research community. Based on different physical properties, we categorize the benchmark samples accordingly. Additionally, we classify VDMs into text-to-video (T2V) diffusion models and image-to-video (I2V) diffusion models based on the input conditions.

In Table 1, we present our experimental results on Phy-

Model	S.M.(↑)	F.D.(↑)	Optics(↑)	Magnetism(↑)	Thermodynamics(↑)	Average(↑)
Cogvideo-I2V-5B	30.4	29.8	16.7	13.3	8.5	27.1
SVD-XT	21.9	20.5	6.8	8.4	**17.1**	19.1
LTX-Video-I2V	30.2	29.8	15.9	13.2	8.4	26.8
SG-I2V	34.6	31.2	15.9	13.1	8.4	29.7
Ours	**42.3**	**34.1**	**16.9**	**13.4**	8.8	**34.6**

Table 2. Quantitative results of physically plausible video generaion on Physics-IQ Benchmark. S.M. refers to Solid Mechanics, and F.D. refers to Fluid Dynamics.

GenBench, evaluating different video generation models following its evaluation protocol. The results show that our framework achieves state-of-the-art performance across four different physical phenomena. **Our framework outperforms the best T2V method by an average of 15.3% and the best I2V method by 11.1%.** Specifically, our framework demonstrates significant advantages in the Mechanics, Thermal, and Material domains, outperforming the best I2V method by 5.7%, 17.6%, and 35.8%, respectively. These advantages are particularly evident in these three types of physical phenomena, which involve more substantial changes in motion, volume, or shape. Our framework is better equipped to understand and reason about bounding box sequences to represent these changes effectively.

Similarly, for the Physics-IQ benchmark, we evaluate the performance of different video generation models following its evaluation protocol. **Our framework achieves the best results across four different physical phenomena, with improvements of 22.2% in Solid Mechanics and 9.2% in Fluid Dynamics compared to the second-best models.** These significant improvements demonstrate the effectiveness of our framework in generating physically plausible videos.

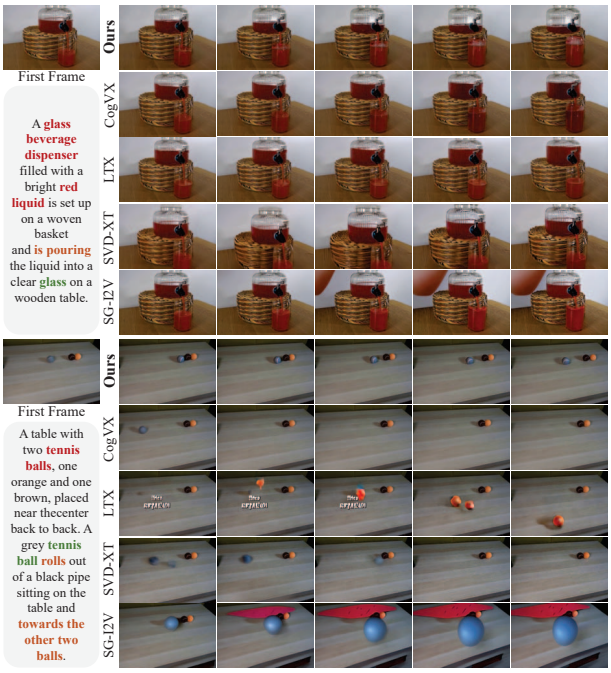

Figure 5. Visual comparisons of physically plausible video generation results from our framework, CogVideoX-I2V-5B, LTX-Video-I2V, SVD-XT and SG-I2V [35] in the Physics-IQ dataset.

4.4. Qualitative Evaluation

Figures 4 and 5 demonstrate a qualitative comparison between our video generation framework and baseline methods. Among all evaluated approaches, our framework consistently produces videos with the highest degree of physical realism. In the ball falling sample in Figure 4, while CogVideoX shows a bouncing effect, artifacts are present in the video; LTX-Video and SVD-XT exhibit motions that do not adhere to the laws of physics. In Figure 5, we analyze two examples from Physics-IQ. In the pouring water example, the baseline methods fail to show the simultaneous decrease in water level of the glass beverage dispenser and the increase in water level of the glass below; in the ball collision example, none of the baseline methods correctly depict the collision of balls. More videos are provided in the supplement.

4.5. Ablation Study

We perform an ablation study to evaluate the contributions of key components in our framework. We design four variants to analyze the effectiveness of different components in our framework.

1. **Ours w/o VLM Planner:** To assess the overall functionality of our framework, we replace the structured noise input of the VDM with random noise to evaluate the effectiveness of the VLM planner.
2. **Ours w/o CI**: Keeping the overall structure unchanged, we remove the in-context information from the VLM.
3. **Ours w/o CoT**: Similarly, while keeping other components unchanged, we remove the CoT reasoning process from the VLM.
4. **Ours w/o CC**: Lastly, we remove both the in-context information and the CoT reasoning process from the VLM planner while maintaining all other components.

Model	S.M.(↑)	F.D.(↑)	Optics(↑)	Magnetism(↑)	Thermodynamics(↑)	Average(↑)
Ours	**42.3**	**34.1**	**16.9**	**13.4**	**8.8**	**34.9**
w/o VLM Planner.	16.3	20.8	13.4	5.8	5.6	16.2
w/o C.I	26.3	28.1	16.9	11.2	8.4	24.3
w/o CoT	21.4	26.9	16.1	8.6	6.9	21.0
w/o C.C	18.7	22.4	14.9	7.2	6.1	18.1

Table 3. Ablation study on VLM, in-context learning and COT. S.M. refers to Solid Mechanics, and F.D. refers to Fluid Dynamics.

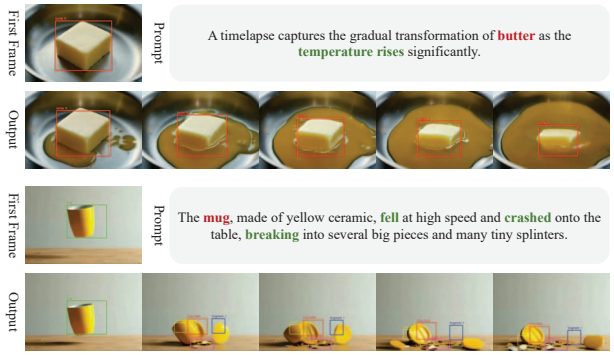

Figure 6. Visualization of generated video with bounding boxes. Coarse motion cue can guide VDM to synthesis physically plausible video.

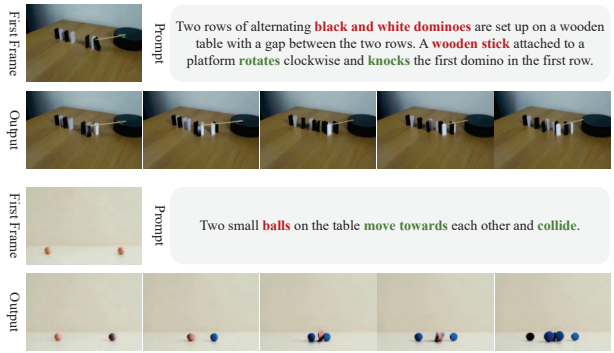

Figure 7. Limitation of VLIPP. Our framework might produces physically implausible videos because it lacks 3D spatial perception and the optical-flow of small objects is prone to noise.

Table 3 presents a quantitative comparison between our full method and these variants. Among all variants, **Ours w/o VLM Planner** shows the most significant performance drop, as removing the planner completely eliminates our ability in understanding the physical laws, leading to nearly random results. Notably, **Ours w/o CoT** exhibits a more pronounced decline compared to **Ours w/o CI**, indicating that the reasoning process in CoT enhances the understanding of physics. While in-context information contributes to the physical reasoning ability of VLM, compared to CoT it is less effective in preventing errors caused by VLM hallucination.

4.6. Limitations

Although VLIPP can generate physically plausible videos, its performance remains constrained by the base model.

Firstly, we are unable to model physical events that cannot be represented by image space bounding box trajectories. For example, phenomena that involve intrinsic state changes of objects such as gas solidification. We adopt 2D bounding boxes because they can be naturally integrated with VDMs and are effective in enhancing the physical plausibility to a certain extent. In fact, they are more flexible than they might appear. As shown in Figure 6, the VLM predicts coarse 2D bounding box trajectories(only four fragments), yet the VDM still produces a realistic shattering pattern without being rigidly constrained by those paths. Moreover, in the butter melting example in Figure 6, the VLM offers only a coarse cue of volumetric change, but the VDM successfully generates the transition from solid to liquid. These abilities significantly improve our flexibility in modeling various physical phenomena.

Secondly, our pipeline lacks 3D spatial perception. It is unable to understand the spatial relationships within the scene, as shown in Figure 7 domino example.

Finally, the optical flow of small objects is prone to noise interference. This will cause our framework to generate ambiguous content, as shown in Figure 7 small ball collision example. With the recent progress in video generation model, we anticipate that our framework will be further improved in generating videos under more challenging physical conditions.

5. Conclusion

Recently, VDMs have achieved great empirical success and are receiving considerable attention in computer vision and computer graphics. However, due to the lack of understanding of physical laws, VDMs are unable to generate physically plausible videos. In this paper, we introduce VLIPP, a novel two-stage physically plausible video generation framework that incorporates physical laws into video diffusion models through vision and language informed physical prior. Our experimental results demonstrate the effectiveness of our method compared to existing approaches.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grant No. 62441231, 62472065, and U23B2010.

References

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

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 3

[3] 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. 4

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024. 2

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6

[6] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. 2

[7] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. *arXiv preprint arXiv:2501.08331*, 2025. 2, 5

[8] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *The Twelfth International Conference on Learning Representations*. OpenReview, 2024. 5

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, 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. 4

[10] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[11] 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. 4

[12] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. *arXiv preprint arXiv:2401.18085*, 2024. 5

[13] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 5

[14] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Real-time video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 6

[15] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[17] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 5

[18] Hao-Yu Hsu, Zhi-Hao Lin, Albert Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically realistic video editing from natural language instructions. *arXiv preprint arXiv:2411.02394*, 2024. 2

[19] Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. Fine-grained controllable video generation via object appearance and context. *arXiv preprint arXiv:2312.02919*, 2023. 3

[20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 3

[21] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 2, 3

[22] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5

[23] Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Guo Zhou, Hua Yao, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models. *arXiv preprint arXiv:2312.00651*, 2023. 3

[24] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024. 2

[25] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 3, 6

[26] 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

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3

[28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 4

[29] 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, 2024. 2, 4

[30] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[31] 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. 4, 5

[32] Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2

[33] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025. 2, 4, 5

[34] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505, 2024. 3

[35] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 6, 7

[36] Andrew Nealen, Matthias Müller, Richard Keiser, Eddy Boxerman, and Mark Carlson. Physically based deformable models in computer graphics. In *Computer graphics forum*, pages 809–836. Wiley Online Library, 2006. 2

[37] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*, pages 111–128. Springer, 2024. 3

[38] Openai. https://openai.com/index/hello-gpt-4o/, 2024. 4

[39] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 3

[40] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024. 3

[41] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4

[42] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4): 1–10, 2013. 2

[43] Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: Biomedical generalist video generation model. *arXiv preprint arXiv:2407.08944*, 2024. 2

[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5

[45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 5

[46] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 3

[47] Xi Wang, Robin Courant, Marc Christie, and Vicky Kalogeiton. Akira: Augmentation kit on rays for optical video generation. *arXiv preprint arXiv:2412.14158*, 2024. 3

[48] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 2

[49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[50] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022. 3, 4

[52] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*, 2024. 3

[53] 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

[54] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 3

[55] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 3

[56] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 2

[57] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024. 3

[58] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. *arXiv preprint arXiv:2412.00596*, 2024. 1, 6

[59] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 2

[60] 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. 5, 6

[61] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3

[62] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023. 2

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[64] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406. Springer, 2024. 2

[65] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 3

[66] 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. *arXiv preprint arXiv:2412.20404*, 2024. 6