Chain of Event-Centric Causal Thought for Physically Plausible Video Generation

Zixuan Wang[1,†] Yixin Hu[1,†] Haolan Wang[1] Feng Chen[2,†] Yan Liu[3]
Wen Li[4] Yinjie Lei[1,*]
[1]Sichuan University [2]The University of Adelaide
[3]Hong Kong Polytechnic University [4]University of Electronic Science and Technology of China
zixuan98@stu.scu.edu.cn, yinjie@scu.edu.cn

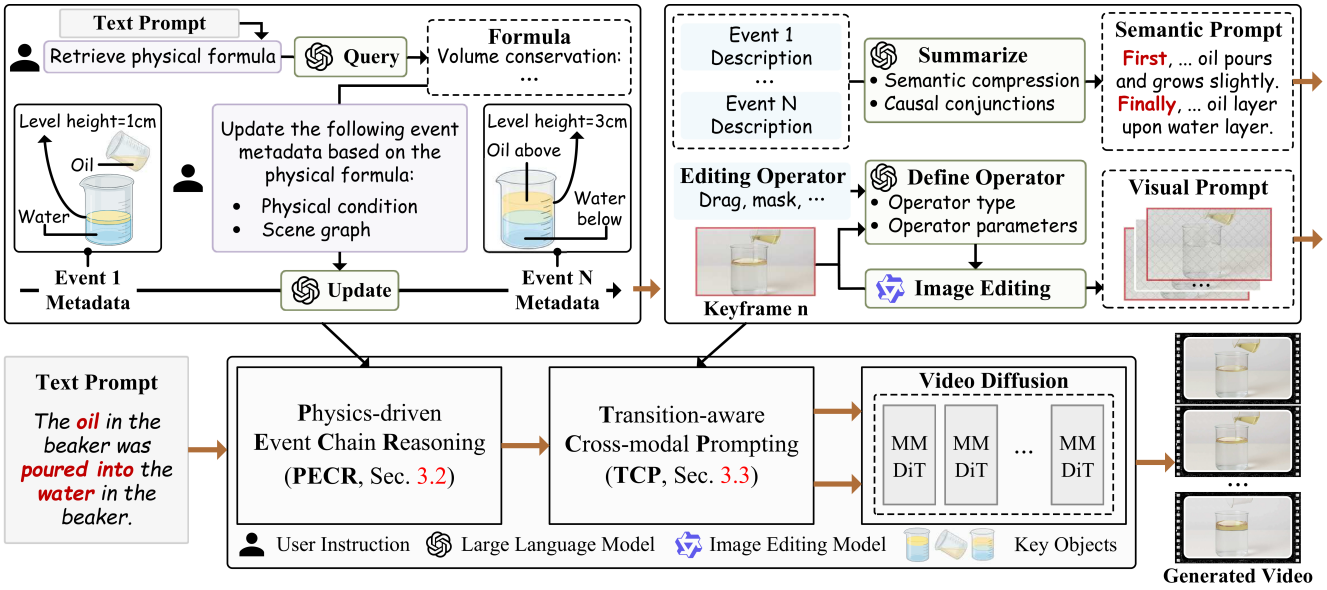

Figure 1. Overview of our physically plausible video generation framework. We firstly decompose complex physical phenomena into a sequence of elementary events guided by physical formulas (Sec. 3.2), and secondly map logically ordered events to a holistic description and a set of keyframes, both of which are causally coherent (Sec. 3.3). Our inferred vision-language prompts enable off-the-shelf diffusion frameworks to generate videos capturing the causal progression of physical phenomena.

Abstract

Physically Plausible Video Generation (PPVG) has emerged as a promising avenue for modeling real-world physical phenomena. PPVG requires an understanding of commonsense knowledge, which remains a challenge for video diffusion models. Current approaches leverage commonsense reasoning capability of large language models to embed physical concepts into prompts. However, generation models often render physical phenomena as a single moment defined by prompts, due to the lack of conditioning mechanisms for modeling causal progression. In this paper, we view PPVG as generating a sequence of causally connected and dynamically evolving events. To realize this paradigm, we design two key modules: (1) Physics-driven Event Chain Reasoning. This module decomposes the physical phenomena described in prompts into multiple elementary event units, leveraging chain-of-thought reasoning. To mitigate causal ambiguity, we embed physical formulas as constraints to impose deterministic causal dependencies during reasoning. (2) Transition-aware Cross-modal Prompting (TCP). To maintain continuity between events, this module transforms causal event units into temporally aligned vision-language prompts. It summarizes discrete event descriptions to obtain causally consistent narratives, while progressively synthesizing vi-

† Equal contribution.
* Corresponding author.

sual keyframes of individual events by interactive editing. Comprehensive experiments on PhyGenBench and VideoPhy benchmarks demonstrate that our framework achieves superior performance in generating physically plausible videos across diverse physical domains. Code is available at https://github.com/ZixuanWang0525/CoECT.

1. Introduction

PPVG has opened up a wide range of real-world applications, including movie production [1], autonomous driving [2], and embodied AI [3]. In recent years, video diffusion models, such as Kling [4] and OpenAI-Sora [5], have demonstrated remarkable capabilities in synthesizing photorealistic scenes from user prompts. However, brief prompts fail to provide the detailed physical laws required for the physically plausible generation. This hinders such generative models from simulating real-world physical phenomena, *e.g.*, fluid dynamics, light refraction, and thermodynamic effects.

Recent PPVG studies [6–8] have augmented user prompts with physical concepts based on Large Language Model (LLM)-assisted reasoning. However, these approaches typically simplify the generated physical phenomena to a single moment defined by static prompts. This challenge arises due to: (1) *Causal Ambiguity*. In the real-world, physical phenomena unfold as causally ordered event units. Unfortunately, embedding a semantic tag to describe such complex phenomena often fails to capture their dynamic nature. This requires a structured decomposition of physical phenomena by causal deterministic reasoning. (2) *Insufficient Physics-consistent Constraints.* Language alone is inherently incapable of conveying the causal continuous between events. Visual cues (*e.g.*, reference videos) can provide observable evidence of event transitions. Even so, visual priors tightly aligned with specified physical phenomena are often hard to obtain.

In this paper, we propose an event-centric physically plausible video generation framework that models physical phenomena as transitions between causally linked events, as shown in Fig. 1. The framework consists of two core modules: (1) We design a Physics-driven Event Chain Reasoning (PECR) module (Sec. 3.2) to decompose physical phenomena into a sequence of fine-grained event units. To mitigate causal ambiguities, we embed computational analysis driven by physical formulas into the reasoning process with scene graphs. This enables the inference of physically realistic events with clear causal relationships. (2) We develop a Transition-aware Cross-modal Prompting (TCP) module (Sec. 3.3) to ensure causal coherence and visual continuity between generated events, through the synergy of semantic and visual prompts. From a semantic perspective, this module compresses multiple event descriptions into a single causally consistent representation using causal conjunctions. On the visual side, this module uses keyframes synthesized by interactive editing as visual prompts, maintaining the smooth transition between events.

We evaluate our framework on PhyGenBench [9] and VideoPhy [10] benchmarks. Our framework significantly outperforms current PPVG approaches on physics-informed metrics across diverse physical domains. Crucially, videos generated by our framework can preserve reasonable chronological order of physical events. Our contributions are summarized as follows:

- We propose an event-centric generation framework that models physically plausible videos as sequences of causally connected and dynamically evolving events.
- To address causal ambiguity, we decompose physical phenomena into causally ordered event units by causal reasoning with deterministic physical constraints.
- To constrain continuous generation between physical events, we synthesize temporally aligned semantic-visual prompts to guide event transitions.
- Comprehensive experiments demonstrate that our framework outperforms existing methods in generating physically realistic and causally coherent videos.

2. Related Works

Physical Plausible Video Generation. To make videos obey physical laws, physics-aware generation has been explored. Several works [11–13] characterize physical phenomena through simulations based on *graphics engines*, which are integrated into diffusion sampling to enhance physical realism. To handle diverse *open-domain physical phenomena*, VideoREPA [14] leverages physical knowledge from foundation models. WISA [15] and PhysHPO [16] guide diffusion models to learn physical phenomena from decomposed principles. VLIPP [17], DiffPhy [8], PAG-SAD [7], and PhyT2V [6] leverage CoT reasoning to design physics-aware prompts. However, physical events unfold as causally ordered processes, while current methods, hindered by lack of causal modeling, often collapse them into a single scene.

Chain-of-Thought in Visual Generation. Recent studies have adapted CoT reasoning [18] from language understanding to visual generation, which are divided into two categories. The first leverages *reasoning before generation* paradigm to augment conditioning signals [19–22]. For example, LayerCraft [21], and GoT [22] enable the generation of multiple objects by reasoning about spatial arrangements. Others embed step-by-step reasoning into the synthesis process through *reasoning during generation* paradigm. Z-Sampling [23] performs diffusion self-reflection, bridging the gap between denoising and inversion. Visual-CoG [24] adopts a chain-of-guidance framework to supervise each generation stage. However, current approaches mainly fo-

cus on semantic and spatial reasoning, neglecting the modeling of deterministic causal relationships.

Dual-Prompt in Video Generation. While natural language defines scene semantic, it often underspecifies detailed geometry and motion. Accordingly, visual cues are introduced to guide video generation, including reference image, spatial layouts, and motion priors. Some works [25–29] employ *reference image* as the appearance prior for generating high-fidelity textures and diverse visual styles. To enhance geometric details, several studies introduce *spatial layouts* during generation. SketchVideo [30] leverages sketches to constrain the contours of objects. DyST-XL [31] and BlobGEN-Vid [32] specify the locations of objects through bounding boxes and blobs, respectively. Given the dynamic nature of videos, recent studies use *motion priors* to capture sophisticated trajectories [33, 34]. However, these approaches primarily constrain individual scenes, lacking the ability to ensure smooth transitions between multiple events.

3. Methodology

3.1. Overall Framework

Given a user-provided linguistic description w of physical phenomenon, our goal is to generate the corresponding physically plausible video $\mathbf{V}$ which characterizes underlying progression of described phenomenon.

$$\Gamma : w \rightarrow \mathbf{V}, \tag{1}$$

where Γ denotes our physics-aware video generation framework. Specifically, our framework is organized as two synergistic modules. In Sec. 3.2, we design a Physics-driven Event Chain Reasoning (PECR) module, which interprets each complex phenomenon described in user-provided description into an ordered collection of physical events. In Sec. 3.3, we develop a Transition-aware Cross-modal Prompting (TCP) module, which bridges the event chain inferred by PECR module to the video generation process. Instead of time-invariant linguistic descriptions and reference images, our TCP module dynamically synthesizes dual-conditions evolving with physical processes.

3.2. Physics-driven Event Chain Reasoning

Physical phenomenon involves the progression of events together with the corresponding changes of the physical parameters. Current studies [6, 8] typically bind a physical phenomenon to an individual object, and simply use a semantic tag to coarsely describe each phenomenon. Unlike such approaches, we conceptualize physical phenomena as a series of causally ordered events, as shown in Fig. 2. Each event can be regarded as a composite unit, encompassing descriptive information of objects' semantic and interactions, as well as measurable physical conditions governed by physics formulas. This enables the characterization of crucial moments in a physical process from both qualitative and quantitative perspectives.

Physics Formula Grounding. To numerically describe physical processes, we perform reasoning over the physics formulas $\mathcal{F}^*$ based on the physical laws $\mathcal{L}$ embedded in the linguistic description w. Specifically, physical laws $\mathcal{L}$ are determined by question answering, where options are defined according to [35]. Formula names associated with physical laws are inferred from the linguistic description. After that, inferred formula names $\mathcal{N}_\mathcal{L}$ are used as queries to retrieve physical formulas $\mathcal{F}^*$ from knowledge bases.

$$\mathcal{F}^* = \mathrm{TopK}_{f \in \mathcal{F}_\mathcal{L}} P(f \mid \mathcal{N}_\mathcal{L}, \mathcal{L}), \tag{2}$$

where $\mathcal{F}_\mathcal{L}$ denotes all formulas in the online knowledge base associated with physical laws $\mathcal{L}$. $P(\cdot)$ is the scoring function over the candidate formulas $\mathcal{F}_\mathcal{L}$. When no direct match of inferred formula names $\mathcal{N}_\mathcal{L}$ is found in $\mathcal{F}_\mathcal{L}$, we regenerate formula names using $\mathcal{F}_\mathcal{L}$. Once the formula is retrieved, the physical parameters required for formula analysis are set by commonesense reasoning.

Physical Phenomena Decomposition. To characterize the scene changes induced by complex physical phenomena, we decompose these phenomena into an ordered sequence of key events $\{\mathcal{E}_t\}_{t=1}^T = \{\{\mathcal{C}_t\}_{t=1}^T, \{\mathcal{G}_t\}_{t=1}^T\}$. Where $\{\mathcal{E}_t\}_{t=1}^T$ denotes the metadata of events, $\{\mathcal{C}_t\}_{t=1}^T$ specifies the physical conditions, $\{\mathcal{G}_t\}_{t=1}^T$ denotes the dynamic scene graph, and T is the event number.

Physical conditions are calculated on the basis of our retrieved physics formulas. Intermediate quantities produced during the analytical calculation provide additional physically meaningful signals. By analyzing whether significant variations occur in physical parameters, the boundaries of physical events can be determined. These boundaries enable the continuous video to be discretized into a sequence of events.

$$\mathcal{C}_t = \left\{ \left(\mathbf{P}_t, \mathcal{F}^*(\mathbf{P}_t)\right) \,\middle|\, \|\mathbf{P}_t - \mathbf{P}_{t-1}\| > \tau_p \right\}, \tag{3}$$

where $\mathbf{P}_t$ denotes the physical parameter vectors of all objects within the t-th event. τ_p is the variation threshold, determining whether the change in physical parameters is sufficient to indicate a new event. In order to ensure physical consistency, the parameters inferred for current event are validated against those of neighboring events by detecting abrupt changes that violate physical continuity. When invalid changes are found, the corresponding parameters and physical contexts are fed back for re-inference.

After that, we update the scene graph based on physical conditions. Given $\mathcal{C}_t$, we update the nodes $\mathcal{V}_t$ and edges $\mathcal{R}_t$ of the scene graph $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{R}_t\}$. For nodes $\mathcal{V}_t$, appearance (*e.g.*, liquid changes color) or semantic label (*e.g.*, burn to ashes) would be updated according to variations in

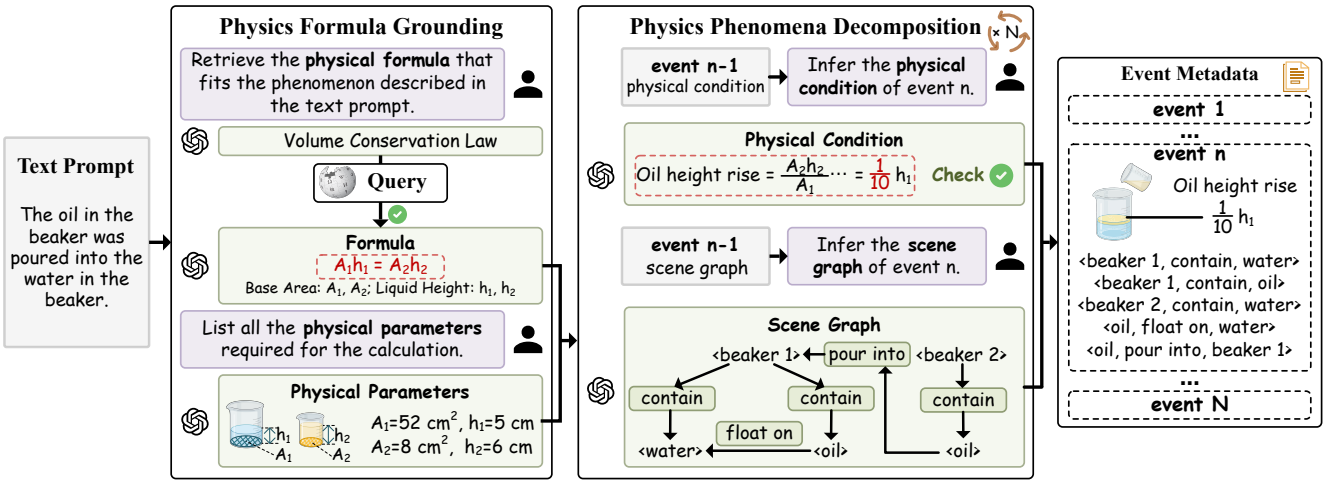

Figure 2. Overview of our PECR module (Sec. 3.2). This module conceptualizes physical phenomena in user-provided descriptions as a series of causally ordered events governed by real-world physical formulas, where each event encompasses semantic descriptions and measurable physical parameters for key objects. This characterizes the underlying scene changes induced by such phenomena.

their physical parameters. Update of edges $\mathcal{R}_t$ is driven by changes of interactions between objects (*e.g.*, a decrease in distance), which requires considering coordinated variations of physical parameters across multiple objects:

$$\mathcal{G}_t = \Phi(\mathcal{G}_{t-1}, \mathcal{C}_t), \tag{4}$$

where $\Phi(\cdot)$ denotes the scene graph update function, which incorporates physical variations captured in $\mathcal{C}_t$ into previous scene graph $\mathcal{G}_{t-1}$ to produce current new graph $\mathcal{G}_t$.

3.3. Transition-aware Cross-modal Prompting

With a sequence of events inferred from the previous stage, the key challenge is to bridge such events to physically realistic video generation. A dual-condition framework [26, 28] serves as a promising solution. However, current frameworks fail to make conditions evolve over time, leading to an inability to capture events' progression. Unlike prior work, we progressively synthesize semantic–visual prompts for each event while preserving seamless temporal progression, as shown in Fig. 3. The semantic prompt serves as a guidance during the denoising and the visual prompt replaces original Gaussian noise to provide physics-aware priors. Generation process is formally defined as:

$$\mathbf{Z}_{\tau_z - 1} = \epsilon_\theta(\mathbf{Z}_{\tau_z}; \mathbf{W}), \tag{5}$$

where $\mathbf{Z}_{\tau_z}$ denotes the visual priors. $\mathbf{W}$ specifies the embedding of linguistic description. ϵ_θ is denoising network of the video diffusion model.

Progressive Narrative Revision. Describing events independently can disrupt holistic coherence of narratives. Therefore, we perform minimal progressive revisions conditioned on the preceding context. Specifically, physical conditions $\mathcal{C}_t$ constrain physically permissible transitions, *e.g.*, rising temperature allows "melting" and excludes "freezing". We employ scene graphs $\mathcal{G}_t$ to preserve object identities and to specify which attributes and relations may change. Let w_t denotes the description of the t-th event, we obtain w_t via:

$$w_t = \mathrm{LLM}\big(w_{t-1} + \Delta(w_{t-1}, \mathcal{C}_t, \mathcal{G}_t)\big), \tag{6}$$

where $\Delta(\cdot)$ denotes an incremental semantic revision guided by $\mathcal{C}_t$ and $\mathcal{G}_t$.

Video diffusion models typically condition on a single sentence. However, concatenating multiple event descriptions produces semantic redundancy. Given this, we merge event descriptions into a positive semantic prompt via semantic condensation and causal connectives. Additionally, we construct a negative description w^*_-. Two descriptions are embedded separately and concatenated as:

$$\mathbf{W} = [\,\psi_{\text{text}}(w^*_+);\ \psi_{\text{text}}(w^*_-)\,], \tag{7}$$

where $\psi_{\text{text}}(\cdot)$ is the text encoder.

Interactive Keyframe Synthesis. While language provides conceptual semantic guidance, physically realistic details remain under-specified, due to the ambiguity of linguistic description. To embed physics-aware details into random Gaussian noise, we synthesize keyframes v_t for each physical event by interactive image editing. These generated keyframes serve as priors to guide video generation process. To be specific, we select the appropriate editing operator from a predefined set (*e.g.*, drag or mask). The change in physical parameters across consecutive conditions acts as a numerical regularizer. It bounds the action space, including dragging magnitude and area of visual change. Practically,

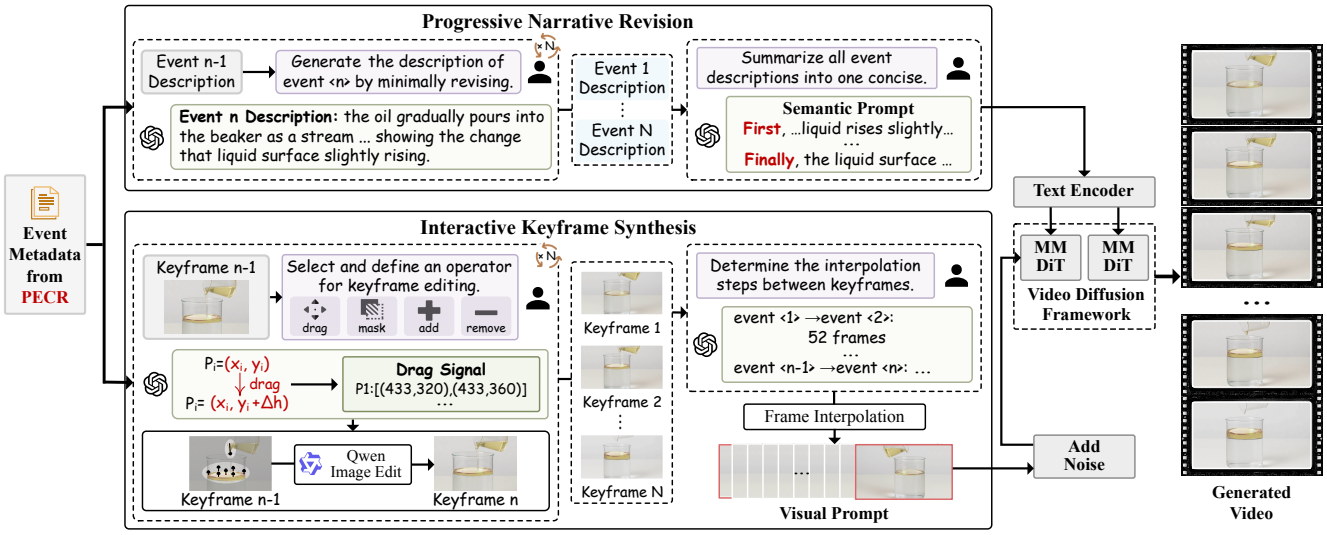

Figure 3. Overview of our TCP module (Sec. 3.3). This module aims to generate semantic-visual prompts for each event based on its meta data. Semantic prompts are inferred by our proposed progressive narrative revision, serving as the guidance during denoising steps. Visual prompts are obtained by our proposed interactive keyframe synthesis, replacing original noise to provide physics-aware priors.

we draw operator cues on a source image to lead the modifications through Qwen-Image-Edit [36]. Each keyframe is obtained as follows:

$$\mathcal{O}_t = \text{LLM}\big((\mathcal{C}_{t-1}, \mathcal{G}_{t-1}) \rightarrow (\mathcal{C}_t, \mathcal{G}_t)\big), \tag{8}$$

$$v_t = \text{Edit}(v_{t-1}; \mathcal{O}_t) \tag{9}$$

where $\mathcal{O}_t$ denotes the editing operator. The keyframe v_1 is directly synthesized from the event description.

To provide a smooth progression, we apply linear interpolation between keyframes. In parallel with inferring $\mathcal{O}_t$, a physically plausible time span d_t for the variation from the $(t-1)$-th to the t-th event is also predicted, which determines how many in-between frames are interpolated between v_{t-1} and v_t. Because video diffusion models often operate in a compressed feature space, we use VAE [37] to encode each keyframe. Based on predicted time span and embedded keyframe features, we perform frame interpolation as follows:

$$\mathbf{z}_{0,t} = \text{INTERP}(\psi_{\text{img}}(v_{t-1}),\ \psi_{\text{img}}(v_t);\ d_t), \tag{10}$$

where $\mathbf{z}_{0,t}$ denotes the interpolated features for variation between two events. $\text{INTERP}(\cdot)$ specifies linear interpolation. $\psi_{\text{img}}(\cdot)$ is VAE encoder. Given a predefined timestep τ_z, we add noise to $[\mathbf{z}_0, \ldots, \mathbf{z}_T]$, serving as priors to denoising process of the video diffusion model.

$$\mathbf{Z}_{\tau_z} = [\mathbf{z}_0, \ldots, \mathbf{z}_T] + \sigma^2_{\tau_z}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{11}$$

where $\sigma^2_{\tau_z}$ denotes the variance of Gaussian noise.

4. Experiments

4.1. Experimental Setups

Datasets. We evaluate on PhyGenBench [9] and VideoPhy [10] datasets. Specifically, PhyGenBench comprises 160 designed linguistic descriptions spanning 27 physical laws across four fundamental domains, namely mechanics, optics, thermal, and material. VideoPhy provides a collection of 688 human-verified linguistic prompts, describing various physical interactions between objects, comprising solid-solid, solid-fluid, and fluid-fluid.

Evaluation Metrics. Following PhyGenBench [9], we use Physical Commonsense Alignment (PCA) as our metric, which indicates video quality by considering key phenomena detection, physics order verification, and overall naturalness evaluation. For VideoPhy [10], we use its provided VideoCon-Physics evaluator to assess Semantic Adherence (SA) and Physical Commonsense (PC). SA evaluates whether a linguistic description is semantically grounded in generated video frames. PC examines whether the depicted actions and object properties conform to real-world physics laws.

Implementation Details. We use CogVideoX 5B [38] as our video generation baseline. Following the official implementation, the video generation model is configured with 161 frames per sample and a resolution of 1360×768. For language reasoning, we employ an open-source GPT-OSS-20B [39] with the default configuration. We use the Qwen-Image family [36] for keyframe generation.

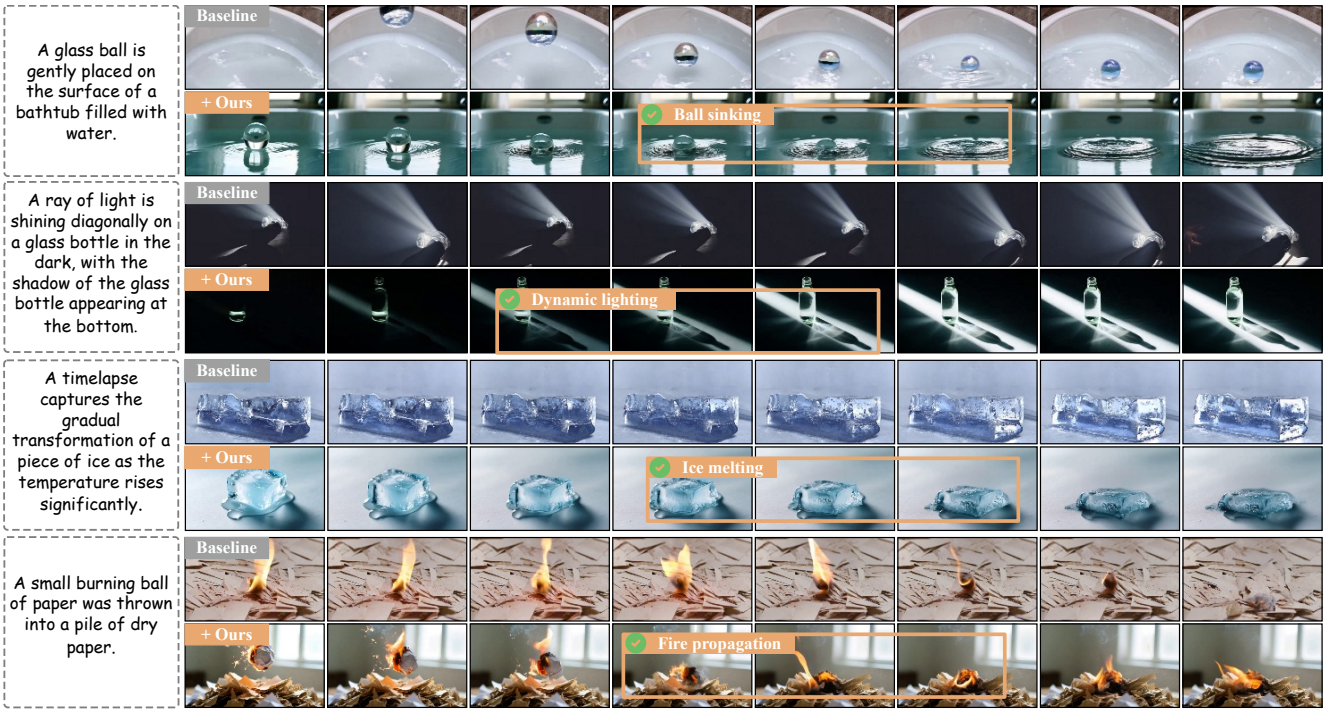

Figure 4. Visualization of physics-aware video generation results across four physical domains. Compared with baseline CogVideo-5B [38], our approach yields causally coherent progressions of physical phenomena, *e.g.*, the glass ball sinks, the bottom shadow extends in the direction of the light, gradual melting of ice, fire spreading through the paper. All prompts are sourced from PhyGenBench [9].

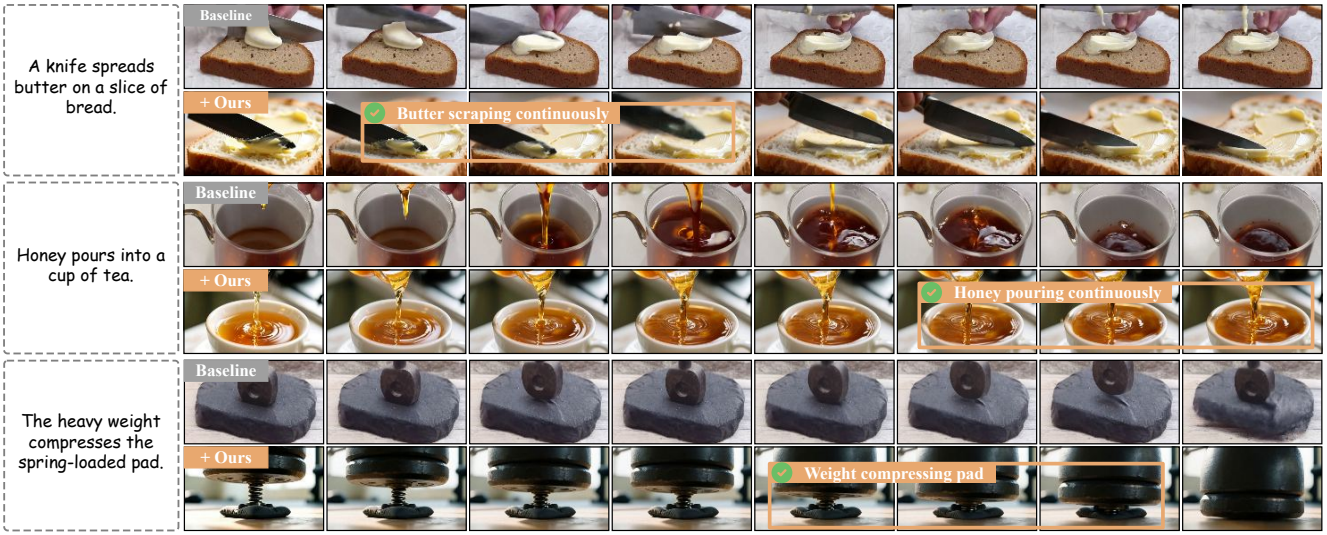

Figure 5. Visualization of physics-aware video generation results across various physical interactions between objects. Compared with baseline CogVideo-5B [38], our approach demonstrates clearer causal progression, *e.g.*, butter spread along the knife movement, continuous honey inflow with a rising level, and monotonic spring compression. All prompts are sourced from VideoPhy [10].

4.2. Evaluation on PhyGenBench

We compare our framework with video foundation models [38, 40–45, 47] and physics-aware video generation models [6–8, 15, 16] on the PhyGenBench [9] benchmark. As shown in Tab. 1, our framework consistently achieves the best overall performance of 0.66, surpassing PhysHPO [16] (previous SOTA) by 8.19% on average. As shown in Fig. 4, compared with the baseline CogVideo-5B [38], our framework achieves visually more realistic generation of gradual

Table 1. Performance comparison on PhyGenBench [9] across four physical domains. The best and second-best results are **highlighted** and underlined, respectively.

Methods	Physical domains (↑)				Avg.
	Mechanics	Optics	Thermal	Material	
Video Foundation Model					
Lavie [40]	0.30	0.44	0.38	0.32	0.36
VideoCrafter v2.0 [41]	-	-	-	-	0.48
Open-Sora v1.2 [42]	0.43	0.50	0.34	0.37	0.44
Vchitect v2.0 [43]	0.41	0.56	0.44	0.37	0.45
Wan [44]	0.36	0.53	0.36	0.33	0.40
Kling [45]	0.45	0.58	0.50	0.40	0.49
Pika [46]	0.35	0.56	0.43	0.39	0.44
Gen-3 [47]	0.45	0.57	0.49	0.51	0.51
Physics-aware Video Generation Model					
WISA [15]	-	-	-	-	0.43
DiffPhy [8]	0.53	0.59	0.58	0.46	0.54
CogVideoX-5B [38]	0.39	0.55	0.40	0.42	0.45
+ PhyT2V [6]	0.45	0.55	0.43	0.53	0.50
+ SGD [7]	0.49	0.58	0.42	0.48	0.49
+ PhysHPO [16]	0.55	0.68	0.50	**0.65**	0.61
+ Ours	**0.67**	**0.72**	**0.65**	0.60	**0.66**

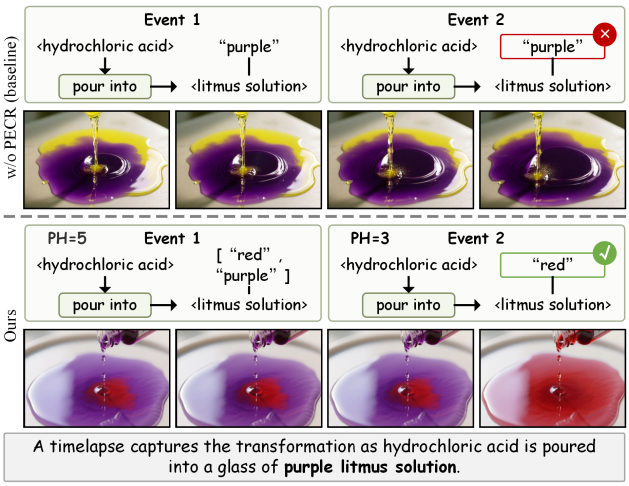

Figure 6. Ablation of the PECR module. Physical-driven CoT reasoning in PECR is crucial for modeling gradual changes in physical conditions and preserving causal consistency across events.

sinking (row 1), light refraction (row 2), realistic melting (row 3), and natural combustion (row 4). These results indicate the ability of our framework to understand physical laws and their corresponding visual dynamics by explicitly considering causally ordered events.

4.3. Evaluation on VideoPhy

We compare our framework with some video generation models [6, 16, 38, 40–42, 46, 48–50] on the VideoPhy [10] benchmark. As shown in Tab. 2, our approach achieves 49.3% scores (SA=1, PC=1) in general, significantly outperforming previous SOTA approach PhysHPO [16] by approximately 3.4%. This confirms the effectiveness of our framework in adhering to the semantic cues and following physical laws by leveraging vision-language prompts inferred via CoT reasoning. As shown in Fig. 5, compared with the baseline CogVideo-5B [38], our approach shows butter spreads along the knife path (row 1), honey pouring as a continuous viscous stream with a steadily rising liquid level (row 2), and the spring shortens monotonically under sustained compression (row 3). These cases emphasize that our framework can capture interactions between objects in diverse physical phases.

4.4. Ablation Studies

We perform ablation studies based on PhyGenBench [9] to systematically analyze the contributions of each block in our proposed PECR and TCP modules.

PFG and PPD blocks in PECR. We analyze the effectiveness of the Physics Formula Grounding (PFG) and Physical Phenomena Decomposition (PPD) blocks in PECR module. When removing the PFG block, physical events are inferred solely from original descriptions; whereas without PPD block, cross-modal conditions are derived only from physical formulas. As shown in Tab. 3, excluding the PFG block causes an average performance drops of about 6% across physical domains, underscoring the necessity of formulas to quantitatively understand physical laws. Similarly, removing the PPD block leads to an average decline of around 11%, indicating its effectiveness in generating realistic physical progressions by decomposing complex processes into logically ordered event chains.

To validate the role of PECR module in our overall framework, we disable its CoT reasoning and verification procedures. In this setup, the module directly infers the scene graph of individual events from original descriptions. As shown in Fig. 6, our framework correctly generates the gradual color transition in the litmus solution from purple to red, whereas the baseline produces a persistently red solution. Upon inspecting inferred scene graphs, we observe that the PECR module updates the color property of the litmus node across events, but the variant leaves it unchanged. This indicates that PECR helps capture the causal progression of events by modeling how physical conditions evolve across events.

PNR and IKS blocks in TCP. We investigate the impact of the Progressive Narrative Revision (PNR) and Interactive Keyframe Synthesis (IRS) blocks in TCP module. As shown in Tab. 3, without PNR block causes a moderate average drop of about 3%, demonstrating its supportive role in improving the continuity and smooth evolution of scenes. Conversely, excluding the IKS block results in a significant average decrease of approximately 17%, which underscores the essential role of explicitly generating dedicated keyframes for each physical phase in anchoring cross-frame dynamics and preserving a physically grounded visual pro-

Table 2. Performance comparisons on VideoPhy [10] across various physical interactions between objects. The best and second-best results are **highlighted** and underlined, respectively.

Methods	Overall (%)			Solid-Solid (%)			Solid-Fluid (%)			Fluid-Fluid (%)		
	SA, PC	SA	PC	SA, PC	SA	PC	SA, PC	SA	PC	SA, PC	SA	PC
Video Foundation Model												
VideoCrafter2 [41]	19.0	48.5	34.6	4.9	31.5	23.8	27.4	57.5	41.8	32.7	69.1	43.6
LaVIE [40]	15.7	48.7	28.0	8.5	37.3	19.0	15.8	52.1	30.8	34.5	69.1	43.6
SVD-T2I2V [48]	11.9	42.4	30.8	4.2	25.9	27.3	17.1	52.7	32.9	18.2	58.2	34.5
OpenSora [42]	4.9	18.0	23.5	1.4	7.7	23.8	7.5	30.1	21.9	7.3	12.7	27.3
Pika [46]	19.7	41.1	36.5	13.6	24.8	36.8	16.3	46.5	27.9	44.0	68.0	58.0
Dream Machine [49]	13.6	61.9	21.8	12.6	50.0	24.3	16.6	68.1	23.6	9.0	76.3	11.0
Lumiere [50]	9.0	38.4	27.9	8.4	26.6	27.3	9.6	47.3	26.0	9.1	45.5	34.5
Physics-aware Video Generation Model												
CogVideoX-5B [38]	39.6	63.3	53	24.4	50.3	43.3	53.1	76.5	59.3	43.6	61.8	**61.8**
+ PhyT2V [6]	40.1	-	-	25.4	-	-	48.6	-	-	55.4	-	-
+ Vanilla DPO [51]	41.3	-	-	28.2	-	-	50.0	-	-	51.8	-	-
+ Ours	**49.3**	**79.5**	**59.4**	**40.6**	**73.4**	**53.8**	**60.0**	**85.6**	**66.7**	54.5	**85.4**	**61.8**

Table 3. Ablation analysis of PECR and TCP modules, including Physics Formula Grounding (PFG) and Physical Phenomena Decomposition (PPD) in PECR, and Progressive Narrative Revision (PNR) and Interactive Keyframe Synthesis (IRS) in TCP.

Variant	Physical domains (↑)				Avg.
	Mechanics	Optics	Thermal	Material	
Ours	**0.67**	**0.72**	**0.65**	**0.60**	**0.66**
Ablations of PECR module					
w/o PFG	0.63	0.69	0.61	0.53	0.62
w/o PPD	0.58	0.67	0.61	0.52	0.59
Ablations of TCP module					
w/o PNR	0.65	0.70	0.64	0.56	0.64
w/o IKS	0.50	0.64	0.58	0.48	0.55

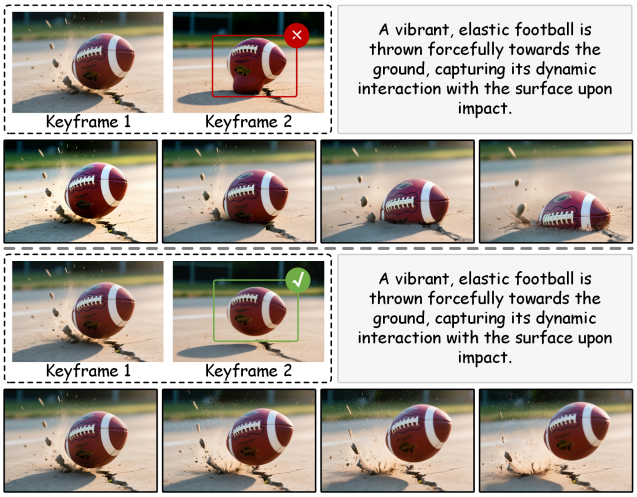

Figure 7. Ablation of visual prompts in TCP module. Physically consistent visual keyframe prompts are crucial, as unrealistic keyframes mislead the generator and lead to structurally implausible content.

gression. We also evaluate the effectiveness of visual prompts in TCP module. As shown in Fig. 7, the rounded shape of the rugby ball is preserved when keyframes provide reasonable visual cues. Conversely, the baseline produces an implausible result in which the ball sinks into the ground rather than resting on the surface. This result shows that the physical realism of visual prompts directly affects the structural fidelity of the generated content.

5. Conclusion

This paper addresses the challenge of generating physics-aware videos characterizing the ordered progression of physical phenomena governed by real-world physical laws. In place of depicting a complex physical phenomenon by simple semantic labels, we decompose each phenomenon into a sequence of causally linked physical events based on standard physical formulas. We infer physical events and translate them into vision-language prompts that correspond to each event. These prompts are then used to condition the video generation process, ensuring alignment with physical dynamics. Comprehensive experiments confirm the effectiveness of our framework in generating physically plausible videos, especially in modeling complex and evolving physical phenomena.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (No. U23B2013, 62276176). This work was also partly supported by the SICHUAN Provincial Natural Science Foundation (No. 2024NSFJQ0023).

References

[1] Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, et al. Generative ai for film creation: A survey of recent advances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6267–6279, 2025. 2

[2] 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

[3] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 2

[4] Kuaishou. Kling. https://klingai.kuaishou.com/, 2024. Accessed: 2024-09-03. 2

[5] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2

[6] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18826–18836, 2025. 2, 3, 6, 7, 8

[7] Yutong Hao, Chen Chen, Ajmal Saeed Mian, Chang Xu, and Daochang Liu. Enhancing physical plausibility in video generation by reasoning the implausibility. *arXiv preprint arXiv:2509.24702*, 2025. 2, 7

[8] Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025. 2, 3, 6, 7

[9] 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. 2, 5, 6, 7

[10] 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. 2, 5, 6, 7, 8

[11] Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, and Frédo Durand. Difftaichi: Differentiable programming for physical simulation. *arXiv preprint arXiv:1910.00935*, 2019. 2

[12] Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019.

[13] Hao-Yu Hsu, Chih-Hao Lin, Albert J Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically realistic video editing from natural language instructions. In *2025 International Conference on 3D Vision (3DV)*, pages 769–780. IEEE, 2025. 2

[14] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025. 2

[15] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025. 2, 6, 7

[16] Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. *arXiv preprint arXiv:2508.10858*, 2025. 2, 6, 7

[17] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. *arXiv preprint arXiv:2503.23368*, 2025. 2

[18] 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 (NeurIPS)*, 35:24824–24837, 2022. 2

[19] Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025. 2

[20] Ziyun Zeng, Junhao Zhang, Wei Li, and Mike Zheng Shou. Draw-in-mind: Learning precise image editing via chain-of-thought imagination. *arXiv preprint arXiv:2509.01986*, 2025.

[21] Yuyao Zhang, Jinghao Li, and Yu-Wing Tai. Layercraft: Enhancing text-to-image generation with cot reasoning and layered object integration. *arXiv preprint arXiv:2504.00010*, 2025. 2

[22] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025. 2

[23] Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. *arXiv preprint arXiv:2412.10891*, 2024. 2

[24] Yaqi Li, Peng Chen, Mingyang Han, Bu Pi, Haoxiang Shi, Runzhou Zhao, Yang Yao, Xuan Zhang, and Jun Song. Visual-cog: Stage-aware reinforcement learning with chain of guidance for text-to-image generation. *arXiv preprint arXiv:2508.18032*, 2025. 2

[25] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah,

Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *European Conference on Computer Vision (ECCV)*, pages 205–224. Springer, 2024. 3

[26] Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. Stiv: Scalable text and image conditioned video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16249–16259, 2025. 4

[27] Bolin Lai, Sangmin Lee, Xu Cao, Xiang Li, and James M Rehg. Incorporating flexible image conditioning into text-to-video diffusion models without training. *arXiv preprint arXiv:2505.20629*, 2025.

[28] Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-Akino, Sharon X Huang, and Tim K Marks. Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9015–9025, 2024. 4

[29] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 3

[30] Feng-Lin Liu, Hongbo Fu, Xintao Wang, Weicai Ye, Pengfei Wan, Di Zhang, and Lin Gao. Sketchvideo: Sketch-based video generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 23379–23390, 2025. 3

[31] Weijie He, Mushui Liu, Yunlong Yu, Zhao Wang, and Chao Wu. Dyst-xl: Dynamic layout planning and content control for compositional text-to-video generation. *arXiv preprint arXiv:2504.15032*, 2025. 3

[32] Weixi Feng, Chao Liu, Sifei Liu, William Yang Wang, Arash Vahdat, and Weili Nie. Blobgen-vid: Compositional text-to-video generation with blob video representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 12989–12998, 2025. 3

[33] Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. Trackgo: A flexible and efficient method for controllable video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 10743–10751, 2025. 3

[34] Xuehai He, Shuohang Wang, Jianwei Yang, Xiaoxia Wu, Yiping Wang, Kuan Wang, Zheng Zhan, Olatunji Ruwase, Yelong Shen, and Xin Eric Wang. Mojito: Motion trajectory and intensity control for video generation. *arXiv preprint arXiv:2412.08948*, 2024. 3

[35] Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, et al. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025. 3

[36] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 5

[37] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[38] 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, 7, 8

[39] OpenAI. Gpt-oss-20b model card, 2025. https://openai.com/index/gpt-oss-model-card/. 5

[40] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025. 6, 7, 8

[41] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 7, 8

[42] 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. 7, 8

[43] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025. 7

[44] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 7

[45] Kling AI. Kling, 2024. https://www.klingai.com/. 6, 7

[46] Pika. Pika, 2024. https://pika.art/. 7, 8

[47] Anastasis Germanidis and Runway Research. Introducing gen-3 alpha: A new frontier for video generation, 2024. https://runwayml.com/research/introducing-gen-3-alpha/. 6, 7

[48] 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. 7, 8

[49] Luma AI. Luma dream machine: Ai video generator, 2024. https://lumalabs.ai/dream-machine. 8

[50] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 7, 8

[51] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, pages 8228–8238, 2024. 8