PhysPO: Physics-Aware Local Preference Optimization for Physically Consistent Video Generation

Zhuoran Yang[1,2] and Yanyong Zhang[1⋆]

[1] University of Science and Technology of China
shanpoyang@mail.ustc.edu.cn
[2] Kuaishou Technology
yanyongz@ustc.edu.cn

Abstract. Despite rapid progress in video diffusion models (VDMs), ensuring semantic adherence and physical commonsense remains a fundamental challenge, as even state-of-the-art systems frequently violate real-world physical laws, such as dynamics, thermodynamics, and optics. While Direct Preference Optimization (DPO) has become a popular post-training strategy for aligning generative models, existing video DPO methods to improve physical commonsense suffer from high computational costs, poorly matched preference pairs, and ambiguous global supervision that fails to localize physical violations. We propose ***PhysPO***, a physics-aware local preference optimization framework for physically consistent video generation. We first introduce *CounterPhyPipe*, a physics-aware counterfactual data construction pipeline that forms video preference pairs with consistent global semantics and local physical violations, enabling meaningful comparisons for preference learning. Then we leverage high-frequency decomposition to introduce a physical mask, which localizes regions where physical state transitions may occur. Finally, we develop *PhysPO*, a physics-aware DPO framework that concentrates supervision on physically relevant regions while enforcing neutrality on background regions. This mechanism reduces gradient noise and mitigates shortcut optimization, encouraging the model to focus on genuine physical discrepancies rather than superficial cues. Extensive experiments demonstrate that PhysPO significantly improves physical commonsense without compromising semantic adherence. Our project page is here[3].

Keywords: Video generation · Physical plausibility · Local Preference optimization

1 Introduction

Text-to-video (T2V) diffusion models have recently achieved remarkable progress, demonstrating strong capabilities in high-resolution and long-form video synthe-

⋆ Corresponding author

[3] https://shanpoyang654.github.io/PhysPO/page.html

sis [11, 15, 23, 35, 50]. However, the physical plausibility of generated videos remains severely constrained. Even state-of-the-art models such as CogVideoX [50], HunyuanVideo [15], and Wan2.2 [35] frequently produce outputs that violate real-world physical laws [3,4], including dynamics (e.g., a metal pendulum swinging against gravity in Wan2.1 [35]), thermodynamics (e.g., burning candles exhibiting abrupt upward flames in CogVideoX [50]), and optics (e.g., straw in water with misplaced bending points that contradict optical refraction laws), as shown in Fig.1. Improving the physics reasoning capability of video generation models can make them closer to real-world simulators, which can benefit a wide range of applications such as video gaming [9,33], autonomous driving [39,42,49], robotics [1, 2, 13], film making [29, 38, 44], *etc.*

Recent works have explored various approaches to better model physical dynamics. *Graphics-based* methods [14] rely on simulation engines to specify physical parameters of simple scenarios such as perfectly elastic collisions and basic rigid body dynamics. Nonetheless, it is impractical to apply them in real-world scenes as the environments are too complex to parameterize. *LLM-based* methods [48,52] leverage large language models to augment prompts with explicit physical laws and phenomena, use the extended prompts as input to simulate the physics by iteratively generating the video. such multi-round refinement substantially increases inference latency. More importantly, these approaches simply rely on LLM-augmented instructions without fundamentally improving the intrinsic physical modeling capability of the video generator itself. To inject physical priors and encourage implicit physics learning, *SFT-based* methods typically fine-tune T2V foundation models on large-scale, high-quality text–video datasets. For example, VideoREPA [54] learns physics through representation alignment with video understanding models. WISA [36] enhances physics awareness by incorporating a mixture of physical experts. However, supervised fine-tuning alone is insufficient to reliably enforce physical consistency [19], as shown in Fig.1 (c). The key reason is that there is no negative training data to provide contrastive signals discouraging physically inconsistent generations. Without such contrastive signals, the model may collapse to superficial correlations and exploit shortcut cues, rather than genuinely internalizing physical laws. The emergence of Direct Preference Optimization (DPO) [30] in video diffusion models [10, 24, 31] offers a promising avenue to overcome the limitations of pure SFT. Recent works, such as PhyGDPO [8] and PhysMaster [26], leverage preference-based learning to further improve physical alignment in generated videos.

However, directly using DPO to improve physical plausibility in video diffusion [8, 10, 24, 26, 31, 37] faces three key challenges. First, preference optimization for text-to-video (T2V) models is computationally expensive and inefficient. For example, PhysCorr [37] formulates a physics-aware DPO objective that requires generating multiple candidates (e.g., eight videos) per prompt, from which the most and least physically plausible samples are selected to form preference pairs. Such repeated sampling leads to heavy computational overhead (e.g., 550 seconds to generate a single 720P video with CogVideoX [50] on an NVIDIA H100). Second, preference pairs constructed from independently sampled noise often lack

"apples-to-apples" correspondence. For instance, methods such as PhyGDPO [8] adopt real videos as preferred samples and randomly generated ones as dispreferred samples. Although such dispreferred samples may violate physical plausibility, independent noise initialization frequently results in different global layouts and motion patterns. Consequently, these pairs rarely form **hard counterfactuals**—videos that preserve identical global scene while differing only in subsequent physical dynamics. Instead of focusing on physical consistency, the supervision signal becomes entangled with variations in content, composition, and motion style, leading to ambiguous and noisy optimization signals. To enable meaningful comparisons, preference pairs should maintain similar backgrounds so that differences arise exclusively from the correctness of physical execution. Such controlled counterfactual supervision enables the model to focus explicitly on physical dynamics, rather than merely preferring globally smoother or visually cleaner videos. Third, existing preference optimization methods [8,10,26,37] treat videos as monolithic entities and overlook local preference cues. However, physical violations are intrinsically localized in space and time, whereas global supervision distributes gradients uniformly across the entire video. This dilutes learning signals in physically relevant regions and may encourage the model to modify irrelevant background content rather than correct the underlying physical errors, resulting in noisy and ineffective optimization.

To overcome these challenges, we first propose a Physics-Aware Video Pair Generation Pipeline, ***CounterPhyPipe***, which generates video preference pairs that exhibit local physical inconsistencies. Instead of generating video pairs from independently sampled noise and relying on human or model-based annotations, we construct preference pairs directly from high-quality real videos. Specifically, each real video is treated as the preferred sample, and we edit it to create a corresponding dispreferred variant. The edited video differs only in localized subsequent physical responses, while preserving global semantics. Specifically, we use a multimodal large language model [32] to extract representative frames that capture the scene layout from real preferred videos, and then propose a counterfactual physical description (e.g., "the ball rolls upward against gravity"). We then employ an image editing model [43] to modify the representative frame according to the counterfactual text and synthesize an erroneous end frame. A first-last-frame-to-video model [35] generates a dispreferred video variation conditioned on the start frame, edited end frame, and counterfactual text, producing hard negatives that maintain identical scene layouts but differ in local physical logic. Then we use CounterPhyPipe to collect a training dataset, *CounterPhy*, containing over 30K physically inconsistent text-to-video preference pairs. This construction of preference pairs eliminates the need—present in conventional DPO — to first generate multiple videos and then annotate them, thereby saving substantial labeling costs. Preference pairs in CounterPhy preserve global scene identity but exhibit local physical violations, which enable meaningful comparisons for subsequent local preference optimization.

Second, to better capture local physics preference while enforcing neutrality in background regions, we formulate a Physics-Aware Local Preference Optimiza-

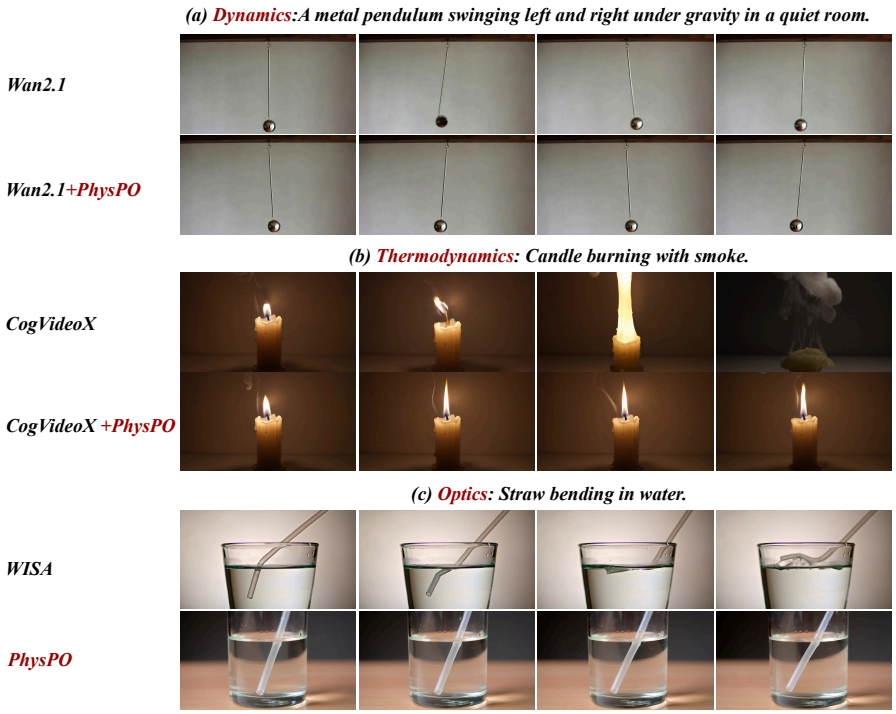

Fig. 1: Text-to-video generation results across three fundamental categories of physics. Videos generated by PhysPO have better semantic adherence and physical commonsense. (a) **Dynamics.** Wan2.1 model generates irregular and non-smooth motion with unstable trajectories that violate basic dynamic constraints. After post-training Wan2.1 with PhysPO, we produce physically consistent pendulum motion, exhibiting smooth periodic oscillations, gradually decaying amplitude due to damping, and gravity-governed trajectories. (b) **Thermodynamics.** CogVideoX exhibits unrealistic abrupt upward flame movement. After post-training CogVideoX with PhysPO, we generate candle-burning videos that follow thermodynamic principles, including gradual wax melting, continuous flame flickering, and upward smoke convection driven by heat. (c) **Optics.** WISA produces geometrically inconsistent distortions, such as misplaced bending points that contradict optical refraction laws. Our model correctly captures refraction-induced visual displacement, where the apparent bending occurs consistently at the air–water interface and varies smoothly with perspective.

tion framework ***PhysPO*** for physically consistent T2V generation. As physical inconsistencies necessarily manifest as abnormal variations in either an object's spatial structure or its temporal evolution, such variations are predominantly reflected in high-frequency components [7, 16], which encode structural boundaries and motion dynamics. Therefore, we perform frequency-domain decomposition to explicitly identify regions where these anomalous variations occur. Concretely, we derive a spatial high-frequency mask via frame-wise 2D FFT followed by high-pass filtering to capture structural details and deformation cues,

thereby locating tangible physical entities. In parallel, we compute a temporal high-frequency mask using 1D FFT and high-pass filtering along the temporal axis at each spatial location to detect motion variations such as velocity changes and temporal discontinuities, highlighting where physical state transitions occur. The union of these two masks yields a unified spatio-temporal region that encompasses both physical entities and their state transitions, forming a candidate space for physics-aware supervision. Building upon this physical mask, we extend the vanilla diffusion DPO objective to a physics-aware formulation that concentrates preference supervision on regions exhibiting potential physical violations. Outside the physical mask, the preferred and dispreferred samples are expected to be semantically equivalent. These areas are therefore treated as *tie regions*, where no preference should be expressed. Accordingly, we introduce a neutrality regularization term that explicitly drives the preference difference toward zero in these non-physical regions. By applying supervision on physically meaningful regions while enforcing neutrality elsewhere, the proposed objective reduces gradient noise, mitigates shortcut optimization, and promotes stable and efficient learning of physical consistency.

In summary, benefiting from our dataset and proposed physics-aware local preference optimization techniques, ***PhysPO*** significantly improves the physical commonsense and semantic adherence of T2V models, and yields better results than the recently best T2V models on some challenging physical categories, as shown in Fig. 1. Our contributions are as follows.

- We propose *CounterPhyPipe*, a physics-aware data construction pipeline that forms video preference pairs with consistent global semantics and local physical violations, enabling meaningful comparisons for subsequent local preference optimization. We use it to collect a training dataset, *CounterPhy*, containing over 30K physically inconsistent text-to-video preference pairs.
- We leverage high-frequency decomposition to introduce a physical mask, which localizes regions where physical state transitions may occur.
- We develop *PhysPO*, a physics-aware DPO framework that concentrates supervision on physically relevant regions while enforcing neutrality on background regions, reducing gradient noise and shortcut optimization while promoting stable alignment with real-world physical laws.

2 Related Works

Physics Plausibility for Video Generation. Although recent text-to-video (T2V) models have achieved strong visual fidelity and motion smoothness [5, 12, 23, 35, 50, 53], multiple benchmarks [3, 4, 25, 28] show that they frequently violate basic physical laws, revealing a gap between perceptual realism and physical validity. Existing efforts address this issue from two perspectives. *Graphics-based* approaches incorporate explicit simulators [20, 27, 46], but remain limited by simulator assumptions and scalability. *LLM-based* methods enhance physics reasoning through prompt refinement or expert routing [36, 47], yet do not fundamentally improve the model's intrinsic dynamics modeling and often depend on

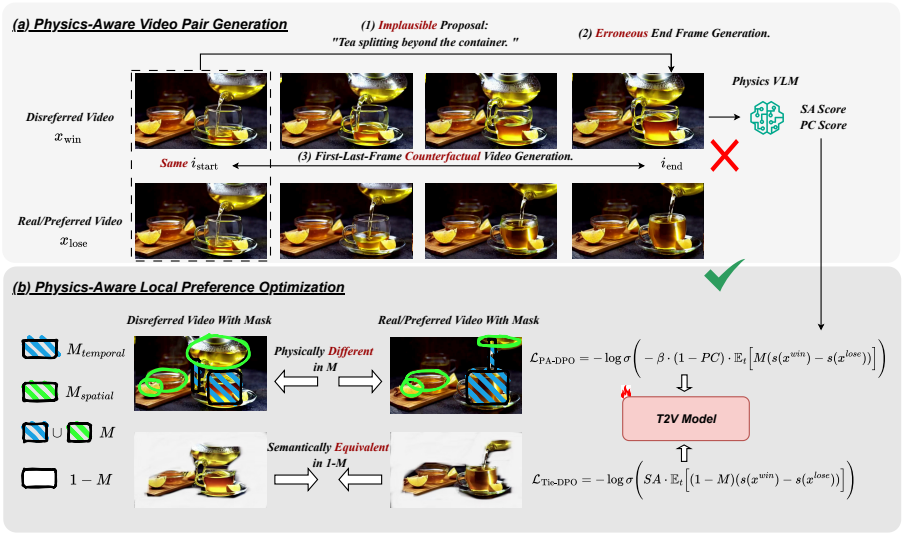

Fig. 2: Overview of our method. (a) CounterPhyPipe: a Physics-Aware Video Pair Generation Pipeline. We regard high-quality real videos as preferred samples and generate dispreferred samples by introducing local physical violations, while keeping the global scene unchanged. (b) PhysPO: a Physics-aware Local Preference Optimization framework. We extract a physical mask to localize regions where physical state transitions may occur. Then we concentrate supervision on physically relevant regions and enforce neutrality on background regions to promote physical plausibility.

curated datasets. Moreover, prior work [19] indicates that supervised fine-tuning alone cannot reliably enforce physical consistency, as models tend to exploit superficial correlations without explicit negative signals. We address this limitation via physics-aware local preference optimization.

Preference Learning for Video Generation. Direct Preference Optimization (DPO) [30] aligns generative models through pairwise supervision without requiring an explicit reward model. Following its extension to diffusion models [34], DPO has been applied to video synthesis [18, 19, 41, 45]. However, existing methods typically rely on global ranking across multiple generated samples, causing preference pairs to reflect coarse perceptual differences rather than localized physical violations. In contrast, we construct preference pairs with controlled local physical discrepancies and optimize alignment within physically relevant regions, enabling fine-grained and physically grounded supervision.

3 Method

We introduce ***PhysPO***, a physics-aware local preference optimization framework that overcomes the limitations of existing DPO-based methods and promotes ***physical commonsense (PC)*** and ***semantic adherence (SA)*** in T2V generation. First, to improve the efficiency and correspondence of constructed video pair, we propose a Physics-Aware Video Pair Generation Pipeline, ***Coun-***

terPhyPipe (Sec. 3.2). We regard high-quality real videos as preferred samples and generate dispreferred samples by introducing local physical violations, while keeping the global scene unchanged. Second, to better capture local physics preference while enforcing neutrality in background regions, we develop a Physics-aware Local Preference Optimization framework, ***PhysPO*** (Sec. 3.3), which explicitly focuses preference learning on physically relevant regions. An overview of PhysPO is illustrated in Fig. 2.

3.1 Background: Video Diffusion and DPO

Rectified-flow diffusion models. Let $\boldsymbol{x} \in \mathbb{R}^{T\times H\times W}$ denote a video sample of length T with spatial dimensions $H \times W$. We follow the rectified flow framework [17, 21], which learns a transport map from the standard normal distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to the distribution of real videos $\boldsymbol{x} \sim p_{\text{data}}$ with a denoiser. The forward diffusion process produces a noisy input $\boldsymbol{x}_t$ at time $t \in [0, 1]$ via a linear interpolation with noise $\boldsymbol{\epsilon}$: $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}$. The denoiser $\mathbf{G}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$, implemented as a neural network parameterized by θ, is trained to reverse this process with the following objective:

$$\min_\theta \mathbb{E}_{t\sim p(t),\boldsymbol{x}\sim p_{\text{data}},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})} \|(\boldsymbol{\epsilon}-\boldsymbol{x}) - \mathbf{G}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})\|^2, \tag{1}$$

where $p(t)$ is the distribution of noise levels and $\boldsymbol{c}$ refers to the auxiliary conditioning variable such as text embeddings.

VanillaDPO. In the DPO framework [30, 34], a generative model is trained to align its outputs with human preferences. Typically, these preferences are defined by a dataset $\mathcal{D} = \{(\boldsymbol{c}, \boldsymbol{x}^w, \boldsymbol{x}^l)\}$, where each sample consists of the preferred and dispreferred videos $\{\boldsymbol{x}^w, \boldsymbol{x}^l\}$ per input condition $\boldsymbol{c}$. The Bradley-Terry (BT) model [6] defines pairwise preference using a reward function $r(\boldsymbol{x}, \boldsymbol{c})$ which computes the alignment score between the sample $\boldsymbol{x}$ and the input condition $\boldsymbol{c}$. The corresponding probabilistic preference can be expressed as:

$$p_{\text{BT}}(\boldsymbol{x}^w \succ \boldsymbol{x}^l|\boldsymbol{c}) = \sigma(r(\boldsymbol{x}^w, \boldsymbol{c}) - r(\boldsymbol{x}^l, \boldsymbol{c})), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function.

DPO [30] defines the binary preference optimization as explicitly optimizing binary reward objective $\log\sigma((r(\boldsymbol{x}^w, \boldsymbol{c}) - r(\boldsymbol{x}^l, \boldsymbol{c})))$ in conjunction with a Kullback-Leibler (KL) divergence regularization to control the deviation from a reference model. Diffusion-DPO [34] re-formulated preference optimization framework for diffusion models, assuming the presence of a reference model $\mathbf{G}_{\text{ref}}$, which is further extended to Flow-DPO in VideoAlign [18]. Given a sample $(\boldsymbol{c}, \boldsymbol{x}^w, \boldsymbol{x}^l)$, denoiser $\mathbf{G}_\theta$, and reference model $\mathbf{G}_{\text{ref}}$, we define an *implicit* reward as:

$$\boldsymbol{s}(\boldsymbol{x}^*, \boldsymbol{c}, t, \theta) = \|(\boldsymbol{\epsilon}^* - \boldsymbol{x}^*) - \mathbf{G}_\theta(\boldsymbol{x}^*_t, t, \boldsymbol{c})\|_2^2 - \|(\boldsymbol{\epsilon}^* - \boldsymbol{x}^*) - \mathbf{G}_{\text{ref}}(\boldsymbol{x}^*_t, t, \boldsymbol{c})\|_2^2, \tag{3}$$

where $\boldsymbol{x}^*_t = (1-t)\boldsymbol{x}^* + t\boldsymbol{\epsilon}^*, \boldsymbol{\epsilon}^* \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is a noisy latent for input $\boldsymbol{x}^*$ (either $\boldsymbol{x}^w$ or $\boldsymbol{x}^l$) at time t. With the implicit reward, the VanillaDPO objective is as:

$$\mathcal{L}(\theta) = -\mathbb{E}_{\substack{(\boldsymbol{c},\boldsymbol{x}^w,\boldsymbol{x}^l)\sim\mathcal{D},\\ t\sim p(t),\,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}} \left[\log\sigma\left(-\beta * \left(\boldsymbol{s}(\boldsymbol{x}^w, \boldsymbol{c}, t, \theta) - \boldsymbol{s}(\boldsymbol{x}^l, \boldsymbol{c}, t, \theta)\right)\right)\right]. \tag{4}$$

3.2 Physics-Aware Video Pair Generation.

To form video preference pairs with consistent global semantics and local physical violations, we propose ***CounterPhyPipe***, a physics-aware counterfactual data construction pipeline, enabling meaningful comparisons for subsequent local preference optimization. We use it to collect a training dataset, ***CounterPhy***, containing over 30K physically inconsistent text-to-video preference pairs.

We start from a large collection of real video–caption pairs (x^{real}, c) and regard high-quality real videos as preferred samples x^w. We generate dispreferred samples x^l by introducing local physical violations into real videos, while keeping the global scene unchanged. Overview of CounterPhyPipe is shown in Fig. 2 (a).

Key Frame Extraction. For each real video–caption pair (x^{real}, c), we select a representative keyframe i^{start} that best aligns with the caption and provides a stable starting state for subsequent editing and generation. Specifically, we employ a shared embedding model [51] to compute caption–frame similarity and choose the frame with the highest alignment score.

Implausible Proposal Generation. Conditioned on the keyframe i^{start} and the original caption c, we prompt a multimodal LLM Qwen3-VL [32] to propose an alternative physical proposal c^l that is: *implausible* according to physical laws (e.g., the ball rolls upward against gravity).

Counterfactual Video Generation. For each alternative physics proposal c^l, we employ the image editing model Qwen-Image-Edit [43] to edit the keyframe i^{start} based on the implausible physical proposal c^l, producing an erroneous end frame i^{end} that reflects the resulting counterfactual physical state. Then, we synthesize the dispreferred video variation x^l with a First-Last-Frame-to-Video model Wan2.1-FLF2V [35], conditioned on the start frame i^{start}, edited end frame i^{end}, and implausible physical proposal c^l. This pipeline yields a set of hard negatives that maintain identical scene layouts but differ in local physical logic, enabling meaningful comparisons for subsequent local preference optimization.

Reward-Guided Video Filtering. To mitigate potential semantic drift introduced by first-last-frame-to-video generation, we employ a reward-guided filtering strategy to construct valid dispreferred samples. Specifically, after each video generation attempt, we utilize the physics-aware VLM VideoCon-Physics [3] to assess whether the synthesized dispreferred video x^l faithfully reflects the intended implausible physical proposal c^l while preserving the original scene identity. VideoCon-Physics produces two normalized scores in $[0, 1]$: a *semantic adherence score* SA, measuring consistency with the input physical proposal c^l, and a *physics commonsense score* PC, quantifying physical plausibility. To construct valid dispreferred samples, we retain videos with higher SA and lower PC, ensuring that the video pair maintains identical scene layouts but differs in local physical logic. Owing to the strong controllability of Wan2.1-FLF2V, only nearly 8% of the generated videos are filtered out. Finally, we collect a dataset, ***CounterPhy*** $\hat{\mathcal{D}} = \{(c, x^w, x^l, PC)_i\}_{i=1}^N$, containing over 30K physically inconsistent video preference pairs.

3.3 Physics-Aware Local Preference Optimization.

To better capture local physics preferences,we propose ***PhysPO***, a physics-aware local preference optimization framework for physically consistent video generation. We leverage high-frequency decomposition to introduce a physical mask, which localizes regions where physical state transitions may occur.Building upon the derived physical mask, we extend the vanilla diffusion DPO objective to a mask-guided, physics-aware formulation that concentrates preference supervision on regions exhibiting potential physical violations while enforcing neutrality on background regions. Overview of PhysPO is shown in Fig.2(b).

Frequency-Domain Physics Filtering. Physical implausibility typically manifests as abnormal variations in spatial structure or temporal evolution. Such variations are largely reflected in high-frequency components [7, 16], which encode structural boundaries and motion dynamics. We construct a physics-aware mask by isolating high-frequency components in spatial and temporal domains.

(1) Spatial Physics Mask. Given a video latent tensor $\mathbf{x} \in \mathbb{R}^{T' \times H' \times W'}$, we perform a frame-wise 2D Fourier transform along spatial dimensions to capture spatial boundaries: $\hat{\mathbf{x}}(t,u,v) = \mathcal{F}_{2D}\big(\mathbf{x}(t,h,w)\big)$. We then apply a radial high-pass filter:$\hat{\mathbf{x}}^{\text{hp}}(t,u,v) = H_s(u,v) \cdot \hat{\mathbf{x}}(t,u,v)$, where $H_s(u,v) = \mathbb{I}\left(\sqrt{u^2+v^2} > \tau_s\right), \quad \tau_s = \beta_s f_s^{\max}$. Here $f_s^{\max} = \max_{u,v}\sqrt{u^2+v^2}$ denotes the maximum spatial radial frequency, and $\beta_s \in (0,1)$ is a fixed ratio. The filtered response is transformed back to the spatial domain: $\mathbf{x}^{\text{sp}}(t,h,w) = \mathcal{F}_{2D}^{-1}\big(\hat{\mathbf{x}}^{\text{hp}}(t,u,v)\big)$. We define a frame-wise adaptive threshold: $\delta_s = \mu_s + \alpha_s \sigma_s$, where μ_s and σ_s denote the mean and standard deviation of $|\mathbf{x}^{\text{sp}}(t,h,w)|$ over spatial locations. The spatial mask is: $M_{\text{sp}}(t,h,w) = \mathbb{I}\left(|\mathbf{x}^{\text{sp}}(t,h,w)| > \delta_s\right)$.

(2) Temporal Physics Mask. To capture temporal dynamics, we perform a 1D Fourier transform along the temporal dimension: $\hat{\mathbf{x}}(f,h,w) = \mathcal{F}_{1D}\big(\mathbf{x}(t,h,w)\big)$. A temporal high-pass filter is applied: $\hat{\mathbf{x}}^{\text{hp}}(f,h,w) = H_t(f) \cdot \hat{\mathbf{x}}(f,h,w)$, where $H_t(f) = \mathbb{I}\left(|f| > \tau_t\right), \quad \tau_t = \beta_t f_t^{\max}$. Here $f_t^{\max}$ denotes the maximum temporal frequency, and $\beta_t \in (0,1)$ is a fixed ratio. After inverse transform: $\mathbf{x}^{\text{tp}}(t,h,w) = \mathcal{F}_{1D}^{-1}\big(\hat{\mathbf{x}}^{\text{hp}}(f,h,w)\big)$. We define a global temporal threshold: $\delta_t = \mu_t + \alpha_t \sigma_t$, where μ_t and σ_t denote the mean and standard deviation of $|\mathbf{x}^{\text{tp}}(t,h,w)|$ over all spatio-temporal positions. The temporal mask is: $M_{\text{tp}}(t,h,w) = \mathbb{I}\left(|\mathbf{x}^{\text{tp}}(t,h,w)| > \delta_t\right)$.

The frequency-domain decomposition is performed on x^w and x^l independently. This symmetric design ensures that all potentially physics-relevant regions are included in the candidate supervision space. We obtain spatial and temporal masks for both samples and construct the final physics-aware mask by taking the union of the four: $M = M_{\text{sp}}^w \cup M_{\text{tp}}^w \cup M_{\text{sp}}^l \cup M_{\text{tp}}^l$. As illustrated in Fig. 2 (b), frequency-domain decomposition effectively isolates physically relevant regions, providing a principled foundation for local preference optimization.

Physics-Aware Preference Optimization. We expect the model to fully capture the divergence between positive and negative samples in the physically relevant regions in $\hat{\mathcal{D}} = \{(c, x^w, x^l, M, PC)_i\}_{i=1}^N$. Therefore, we design a method to extend the vanilla diffusion DPO loss into a physics-aware local preference

optimization objective, denoted by $\mathcal{L}_{\text{PA-DPO}}$:

$$\mathcal{L}_{\text{PA-DPO}} = -\mathbb{E}_{d\sim\hat{\mathcal{D}}}\left[\log\sigma\left(-\beta\cdot(1-PC)\cdot\mathbb{E}_t\left[\boldsymbol{s}'(\boldsymbol{x}^w,\boldsymbol{c},t,\theta,M)-\boldsymbol{s}'(\boldsymbol{x}^l,\boldsymbol{c},t,\theta,M)\right]\right)\right], \quad (5)$$

where $d \triangleq (c, x^w, x^l, M, PC)$ represents the data sample from the preference dataset $\hat{\mathcal{D}}$. Here, $(1-PC)$ serves as a measure of physical implausibility, which adaptively modulates the penalty strength according to the severity of the physical violation. $\boldsymbol{s}'(\boldsymbol{x}^*,\boldsymbol{c},t,\theta,M)$ defines the implicit reward of the current model over the reference model in the physically relevant region M:

$$\boldsymbol{s}'_* = \frac{N_M}{||\mathbf{M}||_1}(\,\|\mathbf{M}\odot((\boldsymbol{\epsilon}^*-\boldsymbol{x}^*)-\mathbf{G}_\theta(\boldsymbol{x}^*_t,t,\boldsymbol{c}))\|^2-\|\mathbf{M}\odot((\boldsymbol{\epsilon}^*-\boldsymbol{x}^*)-\mathbf{G}_{\text{ref}}(\boldsymbol{x}^*_t,t,\boldsymbol{c}))\|^2), \quad (6)$$

where N_M indicates the total number of elements in the physical region $\mathbf{M}$.

Tie-Aware Neutrality Regularization. Outside the physical mask M, the preferred and dispreferred samples are expected to be semantically equivalent. To prevent spurious gradients from physically irrelevant regions, we introduce a regularization term that encourages neutrality in the ***tie regions*** $1-M$:

$$\mathcal{L}_{\text{Tie-DPO}} = \mathbb{E}_{d\sim\hat{\mathcal{D}}}\left[\left(\cdot\,SA\cdot\mathbb{E}_t\left[\boldsymbol{s}'(\boldsymbol{x}^w,\boldsymbol{c},t,\theta,1-M)-\boldsymbol{s}'(\boldsymbol{x}^l,\boldsymbol{c},t,\theta,1-M)\right]\right)^2\right], \quad (7)$$

where SA measures semantic adherence and is used as an adaptive weighting factor, modulating the penalty strength according to the degree of semantic consistency between x^w and x^l. This constraint enforces neutrality inside the physically irrelevant regions and further stabilizes local preference optimization.

Hybrid Training Objective. Excessively prioritizing local pairwise physical preferences may lead to overfitting and impair the model's overall capacity to capture global video structure. To address this, we incorporate standard diffusion DPO as regularization term, thereby promoting stable optimization.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{PA-DPO}}\mathcal{L}_{\text{PA-DPO}} + \lambda_{\text{Tie-DPO}}\mathcal{L}_{\text{Tie-DPO}} + \lambda_{\text{DPO}}\mathcal{L}_{\text{DPO}}, \quad (8)$$

where $\mathcal{L}_{\text{DPO}}$ is the standard diffusion DPO loss applied to the full latent (i.e., with $\mathbf{M}\equiv 1$); $\lambda_{\text{PA-DPO}}, \lambda_{\text{Tie-DPO}}, \lambda_{\text{DPO}}$ are coefficients. This enables PhysPO to learn local physics alignment while preserving global capabilities of base model.

4 Experiments

4.1 Setups

Dataset. We begin by uniformly sampling 40K high-quality text–video pairs from the WISA-80K dataset [36], which comprises 80,000 human-curated videos depicting 17 fundamental physical laws across diverse domains such as dynamics, thermodynamics, and optics. These samples exhibit rich physical interactions and phenomena suitable for physics-aware learning. Following our *CounterPhyPipe*, we construct controlled preference pairs by preserving consistent global semantics while introducing local physical violations. From the filtered subset, we

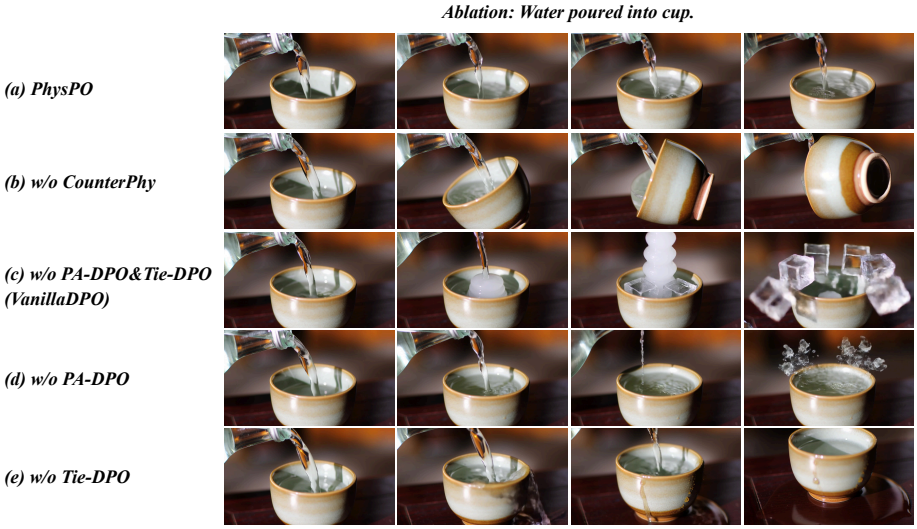

Fig. 3: Qualitative results for ablation study. (a) Our PhysPO generates a video of water being poured into a cup, where the fluid motion follows physical laws. (b) w/o CounterPhy: Lacking physics commonsense, the cup tips over but the water does not spill. (c) w/o PA-DPO and Tie-DPO: Lacking physics commonsense and semantic adherence, the poured water turns into ice cubes. (d) w/o PA-DPO: Lacking physics commonsense, the water spurts abruptly at the surface. (e) w/o Tie-DPO: Lacking semantic adherence, the water is poured outside the cup.

derive 30K valid video preference pairs, denoted as *CounterPhy*, which serve as the training data for physics-aware local preference optimization.

Implementation Details. We implement our method based on Wan2.1-1.3B [35] and CogVideoX-5B [50]. For Wan2.1-1.3B, we finetune for a total of 20K steps at a batch size of 4 on 8 A100 GPUs for 4 days on 30k CounterPhy videos. The training and inference spatial resolution of the video is 480×720. We set $\tau = 3$, $N = 100$, $\alpha_{\min} = 0.5$, $k_\gamma = 2.0$, $b_\gamma = 0.4$, $\lambda = 0.6$, $k_\alpha = 5.0$, $b_\alpha = 0.5$. More implementation details about CogVideoX-5B are shown in the Appendix.

Baselines and Comparisons We evaluate our method on two benchmarks VideoPhy [3] and PhyGenBench [25], to assess the physical commonsense and semantic adherence of generated videos. Our model is benchmarked against its Wan2.1-1.3B [50] and CogVideoX-5B [50] baseline, other leading T2V models (VideoCrafter2 [12], LaVie [40], DreamMachine [22], Cosmos-Diffusion [1], HunyuanVideo [15]), and physics-aware methods like VideoREPA [54], PhyT2V [47], PhyGDPO [8], PhysMaster [26], and WISA [36].

4.2 Quantitative Comparisons

Results in Tab. 1 and Tab. 2 demonstrate that PhysPO achieves SOTA performance across all three interaction types on VideoPhy [3] and PhyGenBench [25].

Table 1: Results of Videophy Benchmark. Semantic Adherence (SA) measures video-text alignment and fidelity. Physical Commonsense (PC) measures whether generated videos follow physics laws. We mark the highest score using blue color.

Methods	Solid-Solid		Solid-Fluid		Fluid-Fluid		Overall		Inference Time
	SA↑	PC↑	SA↑	PC↑	SA↑	PC↑	SA↑	PC↑	(second)
VideoCrafter2	50.4	32.2	50.7	27.4	48.1	29.1	50.3	29.7	-
DreamMachine	55.1	21.7	59.6	23.3	58.2	18.2	57.5	21.8	-
LaVIE	40.8	18.3	48.6	37.0	69.1	50.9	48.7	31.5	-
HunyuanVideo	55.2	16.1	67.1	30.1	54.5	54.5	60.2	28.2	1080
Cosmos-Diffusion-7B	-	-	-	-	-	-	57.00	18.00	600
PhyT2V	-	-	-	-	-	-	61.0	37.0	1800
WISA	-	-	-	-	-	-	67.0	38.0	220
PhysMaster	-	-	-	-	-	-	67.0	40.0	26
Wan2.1-T2V-1.3B	34.9	10.0	64.3	25.4	43.5	48.1	49.0	24.0	180
PhysPO-Wan	55.4	26.0	80.1	44.7	62.0	81.5	67.0(+36.7%)	42.6(+77.5%)	180
CogVideoX-5B	53.1	18.2	75.3	32.9	56.4	61.8	63.1	31.4	170
VideoREPA-CogVideoX	58.0	28.0	82.9	39.0	80.0	74.5	72.1(+14.3%)	40.1(+27.7%)	170
PhysPO-CogVideoX	62.3	31.8	82.5	52.2	87.0	88.9	74.7(+18.4%)	49.4(+57.3%)	170

Table 2: Results of PhyGenBench Benchmark.

Methods	SA↑	PC↑
VideoCrafter	27.0	20.0
OpenSora	23.0	21.0
Cosmos	43.0	14.0
PhyT2V	38.0	42.0
WISA	40.0	43.0
Wan2.1-T2V-1.3B	30.0	22.0
PhysPO-Wan	43.0(+43.4%)	46.0(+109.1%)
CogVideoX-5B	39.0	41.0
PhysPO-CogVideoX	52.0(+33.3%)	64.0(+56.1%)

Physical Commonsense. (1) Comparison with SOTA models. PhysPO gets highest PC score of 49.4 in VideoPhy in Tab. 1, consistently outperforms leading text-to-video models, including VideoCrafter2 [12], LaVie [40], DreamMachine [22], Cosmos-Diffusion [1], and HunyuanVideo [15], as well as physics-oriented approaches such as PhyT2V [47], PhysMaster [26], and WISA [36]. These results validate the effectiveness of our physics-aware local preference optimization, which concentrates supervision on physically relevant regions and promotes stable alignment with real-world physical laws. **(2) Comparison with baseline.** In VideoPhy in Tab.1, starting from the same backbone (CogVideoX-5B), applying PhysPO improves the overall PC score by 57.3%, with gains of 74.7% on Solid–Solid, 58.7% on Solid–Fluid, and 43.9% on Fluid–Fluid interactions. Notably, PhysPO-CogVideoX also surpasses VideoREPA-CogVideoX built on the same base model, further demonstrating its superior ability to enhance physical consistency.

Semantic Adherence. We further evaluate semantic adherence on VideoPhy and PhyGenBench in Tab. 1 and Tab. 2. **(1) Comparison with SOTA models.** PhysPO gets highest SA score of 52.0 in PhyGenBench in Tab. 2, achieving superior SA performance compared to both leading T2V models and existing physics-aware methods, indicating that improvements in physical realism do not come at the expense of semantic alignment. **(2) Comparison with baseline.** When applied to Wan2.1-T2V-1.3B, PhysPO-Wan improves SA score by **43.4%** in PhyGenBench (Tab. 2). This substantial gain demonstrates that our frame-

work enhances physical law adherence while preserving semantic fidelity. Our tie-aware neutrality regularization enforces semantic equivalence between preferred and dispreferred samples on physically irrelevant regions, ensuring stable semantic alignment during optimization.

4.3 Qualitative Comparisons

Fig. 1 presents qualitative comparisons between PhysPO and representative baselines, including Wan2.1-T2V-1.3B, CogVideoX-5B, and WISA.

Physical Commonsense. Our method consistently produces videos with more physically plausible dynamics. In the *"metal pendulum swinging"* scenario in Fig.1 (a), Wan2.1 frequently generates irregular and non-smooth motion with unstable trajectories that violate basic dynamic constraints. In contrast, after post-training Wan2.1 with PhysPO, we produce physically consistent pendulum motion, exhibiting smooth periodic oscillations, gradually decaying amplitude due to damping, and gravity-governed trajectories. Similarly, in the *"candle burning"* example in Fig.1 (c), WISA produces geometrically inconsistent distortions, such as misplaced bending points that contradict optical refraction laws. While PhysPO correctly captures refraction-induced visual displacement, the apparent bending occurs consistently at the air–water interface and varies smoothly with perspective. These results demonstrate that concentrating supervision on physically relevant regions enables more faithful modeling of object interactions and promotes stable alignment with real-world physical laws.

Semantic Adherence. As shown in Fig. 1 (b), compared to CogVideoX, PhysPO better preserves semantic alignment with the input prompts. The generated candle-burning video accurately depicts the text scene while strictly adhering to thermodynamic principles, indicating that improvements in physical commonsense do not compromise semantic adherence. This validates that our framework enhances adherence to physical laws while maintaining semantic consistency.

Table 3: Ablation study results on Wan2.1-T2V-1.3B on PhyGenBench.

Settings	SA↑	PC↑
PhysPO	43.0	46.0
w/o CounterPhy	35.0(-18.6%)	38.4 (-16.5%)
w/o PA-DPO+Tie-DPO(VanillaDPO)	33.2 (-22.8%)	26.0 (-43.5%)
- w/o PA-DPO	39.1 (-9.1%)	29.4 (-36.1%)
- w/o Tie-DPO	34.6 (-19.5%)	40.2 (-12.6%)

4.4 Ablation Studies

We conduct ablation studies on Wan2.1-1.3B on PhyGenBench benchmark to analyze the contributions of our proposed dataset construction pipeline (CounterPhy) and optimization framework (PhysPO) in Tab. 3 and Fig.3.

CounterPhy Dataset. To evaluate the effectiveness of CounterPhy, we replace our counterfactual dispreferred samples with randomly generated dispreferred

videos, while keeping the same training set size and retaining the full PhysPO optimization. Training without CounterPhy consistently decreases both PC and SA score. In particular, PC score decreases by over 16.5%. These results highlight the importance of CounterPhyPipe. By forming preference pairs with global consistent semantics but local physical violations, CounterPhy enables meaningful comparisons and provides informative supervision signals for subsequent local preference optimization. We further present a qualitative ablation analysis in Fig. 3(b). When trained without CounterPhy, the cup tips over, but the water does not spill, violating physical commonsense. This demonstrates that CounterPhy effectively guides the model toward physically coherent state transitions. **PhysPO Optimization. (1) VanillaDPO.** For fair comparison, we fine-tune Wan2.1-1.3B using the same CounterPhy dataset and identical training settings, but we replace our objective with the standard vanilla DPO loss (i.e., without PO-DPO and Tie-DPO). Vanilla DPO consistently underperforms PhysPO across all evaluation tracks in Tab. 3. The performance gap is particularly pronounced in PC, where removing PhysPO leads to an approximately **43.5%** decline. This substantial drop highlights the importance of our physics-aware optimization strategy in effectively modeling physically consistent interactions. Fig. 3(b) further illustrates that applying only vanilla DPO leads to videos with weaker semantic adherence and physical plausibility, in which poured water is unnaturally transformed into ice cubes. **(2) PhysPO w/o PA-DPO.** We further ablate the proposed PA-DPO and Tie-DPO modules individually. Removing PA-DPO leads to substantial performance degradation in PC score. In Tab. 3, PC score drops dramatically (up to 36.1% relative decline), underscoring the critical role of physics-oriented masking in improving physical plausibility. **(3) PhysPO w/o Tie-DPO.** Removing Tie-DPO also results in notable performance decline, particularly in SA score, with up to 19.5% relative reduction. This validates the effectiveness of our tie-aware neutrality regularization, which enforces semantic equivalence between preferred and dispreferred samples. By suppressing spurious gradients from physically irrelevant regions, Tie-DPO ensures stable semantic alignment during optimization while preserving improvements in physical realism.

5 Conclusion

We address the long-standing issue of physical implausibility in T2V models by introducing ***PhysPO***, a physics-aware local preference optimization framework. We first propose *CounterPhyPipe*, a counterfactual data construction pipeline that generates preference pairs with consistent global semantics but local physical violations. Using this pipeline, we build *CounterPhy*, a dataset containing over 30K physically inconsistent T2V preference pairs. Based on this data, we develop *PhysPO*, which focuses supervision on physically relevant regions while enforcing neutrality on tie-regions.Experiments show that our framework significantly improves physical commonsense while preserving semantic adherence, highlighting the effectiveness of our method.

ACKNOWLEDGMENTS This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM113) and the National Natural Science Foundation of China (No. 62332016).

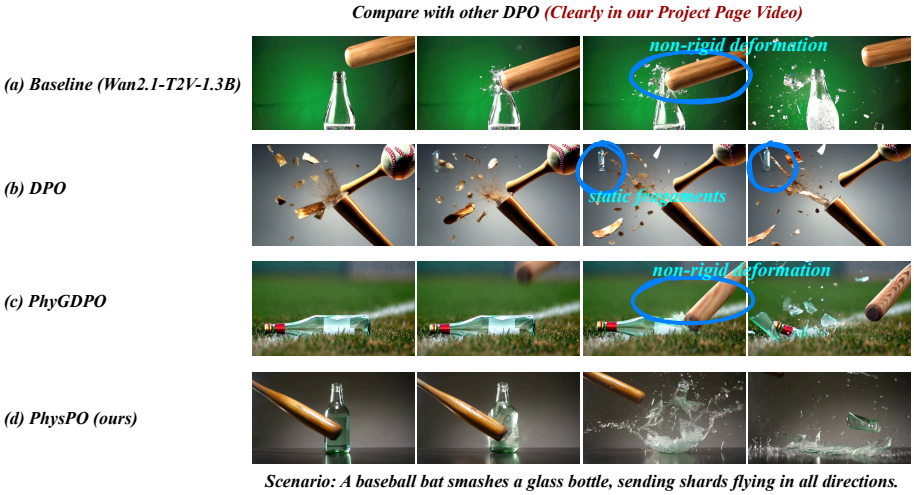

Fig. 4: Qualitative results for comparison with other DPO methods. (a) Baseline (Wan2.1): the glass bottle and baseball bat both undergo non-rigid deformation and develops flexible bending, which does not conform to physical laws. (b) VideoDPO: poor semantic adherence, failing to faithfully represent the shape of the glass bottle. Some fragments (e.g., the transparent shard in the upper-left corner) remain unnaturally static, violating basic gravitational dynamics. (c) PhyGDPO: the baseball bat exhibits non-rigid deformation and flexible bending, violating the physical rigidity expected for such objects. (d) PhysPO (ours): the bat strikes the bottle, leading to visible crack propagation and realistic glass fragmentation.

A More Results

A.1 Comparison with other DPO Methods.

Based on Wan2.1-T2V-1.3B, we compare results after post-training using VideoDPO [19], PhyGDPO [8], and our PhysPO on VideoPhy2.

Table 4: Comparison with other DPO Methods on VideoPhy2 on Wan2.1-T2V-1.3B.

Method	Hard	Activity	Interaction	Overall
Baseline(Wan2.1-T2V-1.3B)	0.0118	0.1136	0.1447	0.1232
VideoDPO [19]	0.0278(+135.6%)	0.1262	0.1509	0.1305
PhyGDPO [8]	0.0444(+276.3%)	0.1357	0.1635	0.1407
PhysPO (Ours)	0.0532(+350.8%)	0.1388	0.1726	0.1470

Quantitative Comparisons For fair comparison with two SOTA DPO methods: VideoDPO and PhyGDPO, we use them to finetune Wan2.1-T2V-1.3B with the same settings. As reported in Tab. 4, PhysPO surpasses PhyGDPO and VideoDPO across all tracks by large margins, especially on the hard action track, where it yields 350.8% improvement over baseline.

Qualitative Comparisons. We present a visual comparison on the challenging collision scenario in Fig. 4. Baseline model Wan2.1 produces a video that shows

the baseball bat undergoes non-rigid deformation and develops flexible bending, which does not conform to physical laws. VideoDPO produces results with poor semantic adherence, failing to faithfully represent the shape of the glass bottle. Moreover, some fragments (e.g., the transparent shard in the upper-left corner) remain unnaturally static, violating basic gravitational dynamics. PhyGDPO generates unrealistic dynamics in which the baseball bat exhibits non-rigid deformation and flexible bending, violating the physical rigidity expected for such objects. In contrast, PhysPO generates physically plausible collision dynamics: the bat strikes the bottle, leading to visible crack propagation and realistic glass fragmentation. The resulting shards scatter outward due to momentum transfer and subsequently fall under gravity, producing coherent and physically consistent motion. Overall, PhysPO maintains stable object structures and realistic force interactions, demonstrating its advantage in modeling complex physical events.

B Method Details

B.1 Reward-Guided Video Filtering Threshold

To mitigate potential semantic drift introduced by first-last-frame-to-video generation, we employ a reward-guided filtering strategy to construct valid dispreferred samples. Specifically, after each video generation attempt, we utilize the physics-aware VLM VideoCon-Physics [3] to assess whether the synthesized dispreferred video x^l faithfully reflects the intended implausible physical proposal c^l while preserving the original scene identity. VideoCon-Physics produces two normalized scores in $[0, 1]$: a *semantic adherence score* SA, measuring consistency with the input physical proposal c^l, and a *physics commonsense score* PC, quantifying physical plausibility.

To construct valid dispreferred samples, we retain videos with higher SA and lower PC, ensuring that the video pair maintains identical scene layouts but differs in local physical logic. The thresholds τ_{sa} and τ_{pc} are determined based on the score distribution over a validation subset. Samples that do not meet these criteria are directly discarded without further refinement. Owing to the strong controllability of Wan2.1-FLF2V, only nearly 8% of the generated videos are filtered out. This high acceptance rate enables an efficient one-pass quality control mechanism with negligible computational overhead, preserving the efficiency advantage of our data construction pipeline. Finally, we collect a dataset, ***CounterPhy*** $\hat{\mathcal{D}} = \{(c, x^w, x^l, PC)_i\}_{i=1}^{N}$, containing over 30K physically inconsistent video preference pairs.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)

2. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025)
3. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)
4. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika (1952)
7. Cai, M., Zhang, H., Huang, H., Geng, Q., Li, Y., Huang, G.: Frequency domain image translation: More photo-realistic, better identity-preserving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13930–13940 (2021)
8. Cai, Y., Li, K., Jia, M., Wang, J., Sun, J., Liang, F., Chen, W., Juefei-Xu, F., Wang, C., Thabet, A., Dai, X., Ju, X., Yuille, A., Hou, J.: Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation (2026), https://arxiv.org/abs/2512.24551
9. Che, H., He, X., Liu, Q., Jin, C., Chen, H.: Gamegen-x: Interactive open-world game video generation. In: ICLR (2025)
10. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: SkyReels-V2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
11. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: VideoCrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
12. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
13. Chen, X., Guo, J., He, T., Zhang, C., Zhang, P., Yang, D.C., Zhao, L., Bian, J.: Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. arXiv preprint arXiv:2411.00785 (2024)
14. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. In: ICML (2025)
15. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: HunyuanVideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
16. Koo, G., Yoon, S., Hong, J.W., Yoo, C.D.: Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. In: European Conference on Computer Vision. pp. 363–379. Springer (2024)
17. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)

18. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Qin, W., Xia, M., Wang, X., Liu, X., Yang, F., Wan, P., Zhang, D., Gai, K., Yang, Y., Ouyang, W.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
19. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: CVPR. pp. 8009–8019 (2025)
20. Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision. pp. 360–378. Springer (2024)
21. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
22. Luma AI: Dream machine | ai video generator (2024), https://lumalabs.ai/dream-machine
23. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248 (2025)
24. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-Video-T2V technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248 (2025)
25. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv preprint arXiv:2410.05363 (2024)
26. Miao, T., Dai, J., Liu, J., Tan, J., Zhang, M., Jin, W., Du, Y., Jin, T., Pang, X., Liu, Z., Guo, T., Zhang, Z., Huang, Y., Chen, S., Ye, R., Zhang, Y., Zhang, L., Chen, K., Wang, W., E, W., Chen, S.: Physmaster: Building an autonomous ai physicist for theoretical and computational physics research (2025), https://arxiv.org/abs/2512.19799
27. Montanaro, A., Savant Aira, L., Aiello, E., Valsesia, D., Magli, E.: Motioncraft: Physics-based zero-shot video generation. Advances in Neural Information Processing Systems **37**, 123155–123181 (2024)
28. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models learn physical principles from watching videos? arXiv preprint arXiv:2501.09038 (2025)
29. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
30. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct Preference Optimization: Your language model is secretly a reward model. NeurIPS (2023)
31. Seawead, T., Yang, C., Lin, Z., Zhao, Y., Lin, S., Ma, Z., Guo, H., Chen, H., Qi, L., Wang, S., Cheng, F., Zuo, F., Zeng, X., Yang, Z., Kong, F., Wei, M., Qing, Z., Xiao, F., Hoang, T., Zhang, S., Zhu, P., Zhao, Q., Yan, J., Gui, L., Bi, S., Li, J., Ren, Y., Wang, R., Li, H., Xiao, X., Liu, S., Ling, F., Zhang, H., Wei, H., Kuang, H., Duncan, J., Zhang, J., Zheng, J., Sun, L., Zhang, M., Sun, R., Zhuang, X., Li, X., Xia, X., Chi, X., Peng, Y., Wang, Y., Wang, Y., Zhao, Z., Chen, Z., Song, Z., Yang, Z., Feng, J., Yang, J., Jiang, L.: Seaweed-7B: Cost-effective training of video generation foundation model (2025)
32. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
33. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. In: ICLR (2025)

34. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: CVPR (2024)
35. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
36. Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al.: Wisa: World simulator assistant for physics-aware text-to-video generation. arXiv preprint arXiv:2503.08153 (2025)
37. Wang, P., Wang, W., Li, Q.: Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection (2025), https://arxiv.org/abs/2511.03997
38. Wang, Q., Luo, Y., Shi, X., Jia, X., Lu, H., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In: SIGGRAPH (2025)
39. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV (2024)
40. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. IJCV (2025)
41. Wang, Y., Tan, Z., Wang, J., Yang, X., Jin, C., Li, H.: Lift: Leveraging human feedback for text-to-video model alignment. arXiv preprint arXiv:2412.04814 (2024)
42. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: CVPR (2024)
43. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
44. Wu, W., Liu, M., Zhu, Z., Xia, X., Feng, H., Wang, W., Lin, K.Q., Shen, C., Shou, M.Z.: Moviebench: A hierarchical movie level dataset for long video generation. In: CVPR (2025)
45. Wu, Z., Kag, A., Skorokhodov, I., Menapace, W., Mirzaei, A., Gilitschenski, I., Tulyakov, S., Siarohin, A.: Densedpo: Fine-grained temporal preference optimization for video diffusion models. arXiv preprint arXiv:2506.03517 (2025)
46. Xie, T., Zhao, Y., Jiang, Y., Jiang, C.: Physanimator: Physics-guided generative cartoon animation. arXiv preprint arXiv:2501.16550 (2025)
47. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. arXiv preprint arXiv:2412.00596 (2024)
48. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation (2025), https://arxiv.org/abs/2412.00596
49. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: CVPR (2024)
50. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)

51. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training. In: ICCV (2023)
52. Zhang, K., Xiao, C., Mei, Y., Xu, J., Patel, V.M.: Think before you diffuse: Llms-guided physics-aware video generation. arXiv preprint arXiv:2505.21653 (2025)
53. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
54. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: Videorepa: Learning physics for video generation through relational alignment with foundation models (2025), https://arxiv.org/abs/2505.23656