
Hierarchical Fine-grained Preference Optimization for Physically Plausible Video Generation

Harold Haodong Chen^{1,2,3}, Haojian Huang^{1,2},
Qifeng Chen^{2,3}, Harry Yang^{†2,3}, Ser-Nam Lim^{†3,4}

¹The Hong Kong University of Science and Technology (Guangzhou)

²The Hong Kong University of Science and Technology

³Everlyn AI, ⁴University of Central Florida

[†]Corresponding Author

✉ haroldchen328@gmail.com

Abstract

Recent advancements in video generation have enabled the creation of high-quality, visually compelling videos. However, generating videos that adhere to the laws of physics remains a critical challenge for applications requiring realism and accuracy. In this work, we propose **PhysHPO**, a novel framework for Hierarchical Cross-Modal Direct Preference Optimization, to tackle this challenge by enabling fine-grained preference alignment for physically plausible video generation. PhysHPO optimizes video alignment across four hierarchical granularities: a) *Instance Level*, aligning the overall video content with the input prompt; b) *State Level*, ensuring temporal consistency using boundary frames as anchors; c) *Motion Level*, modeling motion trajectories for realistic dynamics; and d) *Semantic Level*, maintaining logical consistency between narrative and visuals. Recognizing that real-world videos are the best reflections of physical phenomena, we further introduce an automated data selection pipeline to efficiently identify and utilize "good data" from existing large-scale text-video datasets, thereby eliminating the need for costly and time-intensive dataset construction. Extensive experiments on both physics-focused and general capability benchmarks demonstrate that PhysHPO significantly improves physical plausibility and overall video generation quality of advanced models. To the best of our knowledge, this is the first work to explore fine-grained preference alignment and data selection for video generation, paving the way for more realistic and human-preferred video generation paradigms. [PhysHPO Page](#)

1 Introduction

Video generation has recently achieved significant strides in producing high-quality, visually compelling [39, 73, 90, 99, 64], and lengthy [28, 19, 60, 96] videos depicting real-world scenarios. Despite these advancements, generating videos that adhere to the laws of physics remains a challenging and critical research problem. The ability to create physically plausible videos is essential for applications ranging from virtual reality to simulations, where realism and accuracy are paramount.

Current efforts to enhance the physical fidelity of text-to-video (T2V) generation can be broadly categorized into **test-time reflection-based optimization** and **training-time tuning-based optimization**. Test-time reflection-based methods, such as PhyT2V [85], employ a large language model (LLM) to iteratively refine initial T2V prompts. While effective, these methods significantly decrease computational efficiency and are inherently limited by the upper bounds of the model's capabilities in self-correction paradigms. Conversely, recent tuning-based methods (e.g., WISA [75] and SynVideo [98]) focus on traditional supervised fine-tuning (SFT) paradigms. Although effective, SFT relies



Figure 1: PhysHPO significantly improves the physical plausibility of video generation. Text prompts are adopted from VideoPhy [6]. (Top) *Fluid-Fluid*: Honey diffusing into warm milk. (Middle) *Solid-Fluid*: An apple falls into a vat of red wine. (Bottom) *Solid-Solid*: Peeler peels an apple.

heavily on fixed supervisory signals, which show suboptimal effectiveness when targeting specific capability optimization [43, 51, 23].

Emerging post-training techniques like Direct Preference Optimization (DPO) [61] have demonstrated more efficient pair-wise optimization paradigms to discern differences between human-preferred and non-preferred samples, enhancing specific aspects of visual generation models, *e.g.*, safety [63, 50], customization [43], super-resolution [9]. This indicates DPO holds potential for physically plausible video generation, yet remains largely underexplored. However, recent DPO works in video generation [51, 94, 36, 16] predominantly focus on coarse-grained alignment between videos at the instance level, which may result in suboptimal preference alignment [31]. We emphasize that achieving optimal alignment, particularly for physically plausible video generation, necessitates fine-grained preference alignment that goes beyond visual appeal to incorporate detailed modeling.

Building on these insights, we propose a novel framework, *Hierarchical Cross-Modal Direct Preference Optimization* for physically plausible video generation, namely **PhysHPO**. PhysHPO enhances video preference alignment across hierarchical granularities. Specifically, we design four levels of alignment: ① **Instance Level**: Ensuring comprehensive alignment by matching the overall prompt content with the most suitable video. ② **State Level**: Leveraging boundary frames as critical anchors for establishing plausible states. ③ **Motion Level**: Modeling motion through structural information within videos, enabling alignment beyond mere pixel appearance. ④ **Semantic Level**: Ensuring logical consistency between what is described and what is visually portrayed.

Moreover, compared to existing SFT methods [75, 98] for generating physically plausible videos, which consume extensive resources in dataset construction, we argue that the popular *"One-Model-One-Dataset"* paradigm may not be optimal. Unlike tasks like condition-guided (*e.g.*, dancing [103, 30]) or style-focused (*e.g.*, cartoon [82, 19]) video generation, which demand additional data annotations or domain-specific data, real-world videos inherently encapsulate rich physical dynamics, suggesting potential for more efficient data utilization. To this end, we propose a novel automated data selection pipeline to efficiently process existing large-scale text-video datasets, circumventing exhaustive new data collection efforts for physically plausible video generation. Unlike existing data processing pipelines for high-quality video large-scale pre-training (*e.g.*, in Open-Sora [99]), our key idea is to select a subset of "good data" from large, high-quality raw data that closely matches desired

target requirements—intuitively, real-world videos where physical laws are prominently reflected. To the best of our knowledge, no prior work has explored data selection in the domain of video generation. To summarize, this work contributes in threefold:

- We introduce a novel Hierarchical Cross-Modal DPO (**PhysHPO**) framework for video generation, a more fine-grained DPO strategy to enhance alignment between videos, optimizing for physically plausible video generation.
- We advocate leveraging real-world videos rather than constructing datasets from scratch for physically plausible video generation. This approach intuitively reflects physical phenomena and introduces the data selection problem to video generation for the first time.
- Extensive experiments on both physics-focused (*i.e.*, VideoPhy [6], PhyGenBench [54]) and general capability (*i.e.*, VBench [34]) benchmarks demonstrate that PhysHPO significantly improves the physical plausibility and overall video generation capabilities of existing advanced models.

2 Related Work

Physics-aware Video Generation While generating visually compelling videos has advanced, achieving physics plausibility remains challenging, as noted by users and benchmarks [6, 38, 56, 42, 54, 65]. Existing works [40, 55, 52, 88] in image-to-video (I2V) focus on parsing objects from images and estimating their motion by considering physical properties, and Li et al. [41] explores solely object freefall. However, these methods are limited to fixed physical categories or static scenarios, which restricts their generalizability. Recent works [85, 75, 98] aim to enhance the broader physical plausibility of T2V models. PhyT2V [85] introduces an LLM for iteratively prompt refinement during test time. However, the hugely increased inference overhead and inherent performance limitations restrict its effectiveness. Subsequent research has explored traditional SFT to improve model performance. Specifically, WISA [75] constructs a 32K video dataset, and SynVideo [98] uses a computer graphics pipeline to synthesize video data, both requiring substantial manual intervention, which is resource-intensive. Intuitively, real-world videos naturally reflect physical phenomena. Thus, efficiently utilizing existing datasets without unnecessary data inefficiency is an intriguing problem. To this end, we introduce the concept of data selection for video generation for the first time, automating the process to leverage available data resources effectively. While SFT has demonstrated superior performance in pre-training, we propose adopting the DPO post-training paradigm to further model differences between video pairs, enabling deeper exploration of physical information.

Data Selection for "Good Data" Data selection is a pivotal technique for efficiently training models without sacrificing performance [3, 74], as highlighted in both pre-training [7, 70] and post-training [14, 47] stages. Recent studies emphasize that the effectiveness of LMs stems from a combination of large-scale pre-training and smaller, meticulously curated instruction datasets [69, 17, 18, 77, 101]. Various sophisticated approaches have been proposed for LMs, including quality-based [45, 20, 46], diversity-aware [24, 53], complexity considerations [84, 67, 35], alongside simpler heuristics like selecting longer responses and gradient-based coreset selection [80, 95, 59], collectively demonstrating significant benefits in LM training. However, existing methods predominantly target LMs, leaving their potential in video generation largely unexplored. In this work, we pioneer the exploration of data selection strategies specifically for video generation in the post-training phase. We propose a new automated data selection pipeline, explicitly emphasizing *reality*, *physical fidelity*, and *diversity* to enhance the generation of physically plausible videos.

Direct Preference Optimization for Video Generation DPO [61] has emerged as a promising alternative to traditional RLHF [58] for enhancing LMs [31, 27, 49] and image generative models [72, 78] without needing an extra reward model. Recent efforts have applied DPO to video generation. Pioneering work like VideoDPO [51] follows the Diffusion-DPO [72] paradigm, introducing OmniScore for adaptive video scoring, while HuViDPO [36] aligns outputs with human preferences using feedback. OnlineVPO [94] optimizes off-policy with a video-centric model, and MagicID [43] uses DPO for ID-conditioned customization. Cheng et al. [16] employs a discriminator-free approach for direct optimization, and GAPO [102] introduces AnimeReward for anime video generation. Recent foundation models like Seaweed-7B [64], SkyReels-V2 [10] also incorporate DPO in post-training, further highlighting its potential. However, existing DPO works focus solely on coarse-grained alignment at the video instance level, overlooking finer details. In this paper, we aim to enhance DPO's effectiveness, specifically for generating physically plausible videos by modeling fine-grained alignment. To achieve this, we propose a novel *hierarchical cross-modal preference optimization* framework, named PhysHPO, which effectively captures the fine-grained details of videos.

3 Preliminaries

Direct Preference Optimization for Diffusion Models Diffusion-DPO [72] adapts human preference alignment to iterative generation processes, eliminating explicit reward modeling. Given a diffusion model $p_\theta(y|x, t)$ and reference model $p_{\text{ref}}(y|x, t)$, the denoising objective with preference constraints becomes

$$\max_{p_\theta} \mathbb{E}_{t,x,y \sim p_\theta} [r(x, y, t)] - \beta D_{\text{KL}}(p_\theta(y|x, t) \parallel p_{\text{ref}}(y|x, t)), \quad (1)$$

where $r(\cdot)$ denotes the time-dependent reward function, x the input condition, y the generated sample, and t the timestep. DPO establishes a trajectory-level reward mapping through denoising paths:

$$r(x, y, t) = \beta \log \frac{p_\theta(y|x, t)}{p_{\text{ref}}(y|x, t)} + \beta \log Z(x, t), \quad (2)$$

where β controls KL constraint strength, and $Z(x, t)$ the time-dependent partition function.

The preference optimization objective is derived by substituting the reward parameterization and applying negative log-likelihood loss:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \log \sigma(u(x, y_w, y_l, t)), \quad u = \underbrace{\log \frac{p_\theta(y_w|x, t)}{p_{\text{ref}}(y_w|x, t)}}_{\text{Preferred path score}} - \underbrace{\log \frac{p_\theta(y_l|x, t)}{p_{\text{ref}}(y_l|x, t)}}_{\text{Non-preferred path score}}. \quad (3)$$

4 Data Selection: From Physical to Diverse

Existing text-video datasets have already been rigorously screened for quantity and quality [99, 57, 48, 15]. While creating specific datasets for different tasks is popular [93, 33, 19, 92], for physically plausible video generation, real-world videos inherently reflect physical laws. Thus, selecting existing data may be more optimal than constructing new datasets [75, 98]. In this section, we present an initial exploration of the data selection in video generation, focusing specifically on identifying "good data" that effectively reflects physical laws.

4.1 The Data Selection Problem

Video generation models undergo pre-training to learn world knowledge [8, 2]. Fine-tuning these models aligns them with specific goals, analogous to instruction tuning in LMs [91]. Data selection has proven effective for aligning LMs by identifying a small, high-quality dataset [53, 81, 25, 80]. Consider a large data pool $D = \{x_1, x_2, \dots, x_n\}$, where each data point $x_i = (V_i, C_i)$ consists of a video V_i and its corresponding caption C_i . The objective is to select a subset $S^{(m)}$ of size m that maximizes the post-training performance $\mathcal{P}$:

$$S^* = \arg \max \mathcal{P}(S^{(m)}). \quad (4)$$

While there are various potential ways for data selection in a desired domain, our goal is to keep the process *as simple as possible to be practical*, as shown in Figure 2 (Left). We next detail the process.

4.2 Selection for Real-World Physics

Reality Given a large-scale high-quality text-video dataset D (we adopt OpenVidHD-0.4M [57] in this work, which is widely utilized for post-training [12, 87, 68, 13]), the first step is to filter out real-world videos into a real-world data pool D' . Specifically, considering the significant content differences between real-world and virtual videos, we employ a vision-language models (VLMs) (*i.e.*, Qwen2.5-VL [4]) to determine whether a video V_i is a real-world video. We further task DeepSeek-VL2 [79] for a double-check to ensure the accuracy. Table 1 shows the accuracy from a random sample of 1,000 videos from D .

Table 1: Accuracy of two-stage reality selection with 1,000 samples.

Reality Selection	Acc (%)
+ Qwen2.5-VL [4]	99.6
+ DeepSeek-VL2 [79]	100.0

Physical Fidelity To align with previous works, we adopt 17 physical phenomena across three physics categories from WISA [75]: **① Dynamic** (rigid body motion, collision, liquid motion, gas motion, elastic motion, deformation), **② Thermodynamic** (melting, solidification, vaporization, liquefaction, combustion, explosion), and **③ Optic** (reflection, refraction, scattering, interference and diffraction, unnatural light source). Unlike widely used heuristic classification-based data selection methods in LMs that require a target dataset for reference [81, 89, 21], we follow the popular LLM-as-a-Judge paradigm [53, 26, 102] to employ LLMs for automatic data evaluation.

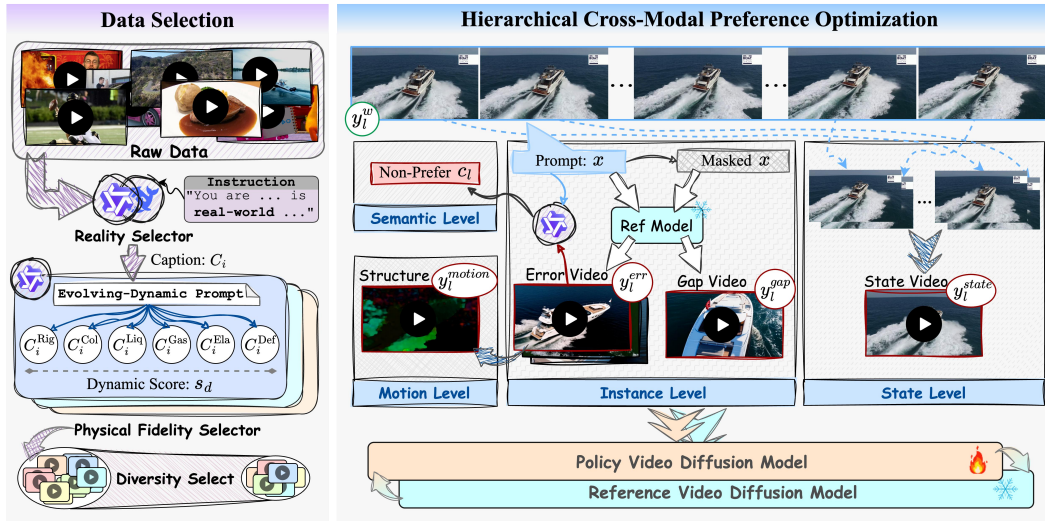


Figure 2: The overview of our proposed (*Left*) data selection and (*Right*) PhysHPO framework.

To minimize consumption, we then perform selections at the caption level. Given a video caption C_i , a naive way is to task an LLM with evaluating whether the sample clearly reflects physical laws by directly assigning a score. However, LLMs might assign similar scores to most samples due to the lack of references [32]. To address this, we propose adopting *in-depth evolving prompting* [84] to augment captions. Unlike previous works [53, 84] that augment samples in a single dimension (e.g., complexity), we explore multiple dimensions. Specifically, as illustrated in Figure 2, we first use the crafted *evolving-dynamic* prompts with an LLM (i.e., Qwen2.5 [86]) to augment the caption, emphasizing each of the seven physical phenomena in the "Dynamic" category. This helps the model distinguish differences in captions more precisely, achieving fine-grained scoring. We then ask the LLM to rank and score these seven samples, obtaining the "Dynamic" scores s_d corresponding to the caption C_i . Similarly, the "Thermodynamic" score s_t and "Optic" score s_o can be obtained. This strategy provides a more nuanced distinction of physical phenomena. Details are in **Appendix §B**.

The total score is calculated as $s = s_d \times s_t \times s_o$. Then all samples in D' are sorted based on s , resulting in the sorted pool $S' = \{x'_1, x'_2, \dots, x'_k\}$, where x'_0 is the sample with the highest score.

4.3 Selection for Diversity

To ensure an advanced generation model can handle varied user prompts, it's desirable for data to maintain maximum diversity within a given budget m . However, real-world data often exhibits redundancy [1]. To this end, we introduce an iterative method to ensure diversity in selected real-world data following Deita [53]. Briefly put, the key idea is to iteratively select samples x'_i from the data pool S' , adding them to the dataset S only if they contribute to its diversity. Formally, we define an indicator function $\mathbb{1}[\mathcal{F}(x'_i, S')]$, which is 1 if the diversity criterion $\mathcal{F}(x'_i, S')$ is met, otherwise 0. Using LLaMA-1 13B [71], captions are encoded into embeddings, and cosine distance d is calculated between each candidate sample and its nearest neighbor in S . A sample contributes to diversity if $d < \tau$, with $\tau = 0.9$.

We compare the performance of different data selection strategies with vanilla DPO [72], along with our PhysHPO, as demonstrated in Figure 3. Several key observations are summarized: **Obs.① Superiority of Data Selection.** Our data (21K), selected for reality and physical fidelity, demonstrates superior quality compared to other strategies like direct score, random selection, or even the manually constructed dataset WISA-32K [75] and unselected raw data (433K). "Our Data (21K)+DPO" outperforms "Direct Scoring (21K)+DPO", "Random Selection (21K)+DPO", and "Raw Data (433K)+DPO", highlighting the effectiveness of our selection. **Obs.② Importance of Data Diversity.** The comparison between "Our

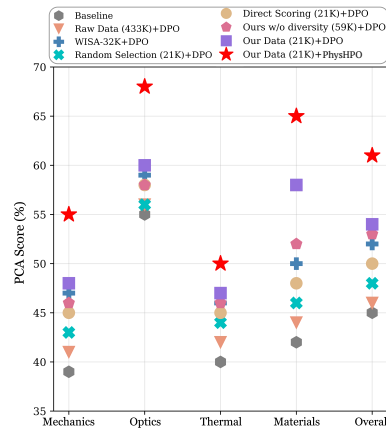


Figure 3: Performance comparison of different data selection strategies with our PhysHPO on PhysGenBench [54].

Data (21K)+DP0" and "Ours w/o diversity (59K)+DP0" underscores the critical role of diversity in data selection. Despite fewer samples, the diverse dataset outperforms the larger one, indicating that diversity enhances model capability. **Obs.③ Strategic Data and Optimization Synergy.** The integration of our selected data with PhysHPO ("Our Data (21K)+PhysHPO") achieves the highest performance, illustrating the powerful synergy between strategic data selection and advanced optimization. More analysis is provided in **Appendix §B**. We next detail our PhysHPO framework.

5 Methodology

To tackle the challenging physically plausible video generation, we propose PhysHPO, which implements hierarchical preference optimization across four levels (Figure 2): (i) Instance-level Overall Preference Optimization, aligning video-level preferences to ensure overall quality (§5.1); (ii) State-level Boundary Preference Optimization, anchoring physical states at the start and end frames for stability (§5.2); (iii) Motion-level Dynamic Preference Optimization, utilizing structural information to accurately model and align motion representation (§5.3); (iv) Semantic-level Consistency Preference Optimization, ensuring fine-grained coherence between the narrative and visuals (§5.4).

5.1 Instance-level Overall Preference Optimization

Unlike VideoDPO [51], which relies solely on its base model for preference pair video data (y_w, y_l) , limiting optimization to the model's capabilities, our approach leverages data selected through our data selection process as the preferred video y_w and its text prompt x . For the non-preferred video y_l , recent works [102, 94] often generate these using the base model based on x , adhering to Eq.(3) as the objective function. However, complex dependencies in non-preferred videos are often overlooked. Typically, video generation faces two main issues: ① Imperfections or errors in physical aspects, *e.g.*, motion; ② Failure to fully express the prompt's information, *e.g.*, missing objects. To mitigate this, we introduce two types of non-preferred videos into the optimization process:

$$\mathcal{L}_{\text{Instance}} = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \log \sigma \left(\beta \log \frac{p_{\theta}(y_w|x, t)}{p_{\text{ref}}(y_w|x, t)} - \beta \log \frac{p_{\theta}(y_l|x, t)}{p_{\text{ref}}(y_l|x, t)} \right), \quad (5)$$

where

$$\log \frac{p_{\theta}(y_l|x, t)}{p_{\text{ref}}(y_l|x, t)} \leftarrow \sum_{i \in \{\text{err}, \text{gap}\}} \beta_i \log \frac{p_{\theta}(y_l^i|x, t)}{p_{\text{ref}}(y_l^i|x, t)}. \quad (6)$$

Here, y_l^{err} denotes the error-prone or imperfect non-preferred video that aligns semantically with the preferred sample. In contrast, y_l^{gap} represents non-preferred videos with semantic differences from the preferred sample. Specifically, for constructing y_l^{err} , we generate three videos per x using the base model and select the one most visually similar by similarity calculation to the preferred video y_w as the rejected video y_l^{err} . For y_l^{gap} , we introduce random masking in prompt x during generation to create semantic differences from y_w .

5.2 State-level Boundary Preference Optimization

While the instance-level alignment captures overall content, fine-grained alignment is essential for generating physically plausible videos, which previous approaches have often overlooked. In video generation, maintaining a consistent physical state from the beginning to the end of a sequence is essential for ensuring realism and coherence. The initial and final frames are particularly critical as they anchor the video's physical narrative. To this end, we propose the state-level boundary alignment to enhance the model's focus on the physical states at the start and end of videos. Specifically, we replace the first and last N frames of the preferred video y_w to construct the state-level non-preferred sample y_l^{state} . The state level objective function can be defined as

$$\mathcal{L}_{\text{State}} \sim \log \sigma \left(u_{\text{State}}(x, y_w, y_l^{\text{state}}, t) \right). \quad (7)$$

5.3 Motion-level Dynamic Preference Optimization

While the visual appearance of videos is crucial for quality, excessive focus on appearance may obscure the model's alignment with physical dynamics. To address this issue and enhance the model's learning of dynamic information such as motion, we propose the extraction of structural information (*e.g.*, optical flow) from both preferred y_w and non-preferred y_l (specifically y_l^{err}) videos. These structural features, commonly used in condition-guided video generation [83, 22, 100], allow for a more explicit representation of the physical motion differences between preferred and non-preferred samples. Following Eq.(7), the motion-level objective function $\mathcal{L}_{\text{Motion}}$ is derived

Table 2: Evaluation on physics-focused benchmarks, *i.e.*, VideoPhy [6], PhyGenBench [54]; and general quality benchmark, *i.e.*, VBench [34]. Vanilla DPO is implemented with our selected data. We **bold** the best results, and "↑" denotes that higher is better.

Method	VideoPhy [6]				PhyGenBench [54]					VBench [34]		
	SS↑	SF↑	FF↑	Over.↑	Mechanics↑	Optics↑	Thermal↑	Materials↑	Overall↑	Total↑	Quality↑	Semantic↑
CogVideoX-2B [90]	12.7	21.9	25.4	18.6	0.38	0.43	0.34	0.39	0.39	81.6	82.5	77.8
+ PhyT2V [85]	14.1	19.9	28.6	18.9	0.45	0.48	0.34	0.50	0.45	82.0	82.9	78.4
+ Vanilla DPO [72]	15.5	19.2	28.6	19.2	0.43	0.50	0.40	0.50	0.46	82.2	83.1	79.0
+ PhysHPO (Ours)	20.4	24.7	42.9	25.9	0.50	0.56	0.47	0.58	0.53	82.5	83.3	79.3
CogVideoX-5B [90]	24.4	53.1	43.6	39.6	0.39	0.55	0.40	0.42	0.45	81.9	83.1	77.3
+ PhyT2V [85]	25.4	48.6	55.4	40.1	0.45	0.55	0.43	0.53	0.50	82.3	83.3	78.3
+ Vanilla DPO [72]	28.2	50.0	51.8	41.3	0.48	0.60	0.47	0.58	0.54	82.4	83.3	78.7
+ PhysHPO (Ours)	32.4	54.1	58.9	45.9	0.55	0.68	0.50	0.65	0.61	82.8	83.7	79.3

from $u_{\text{Motion}}(x, y_w \rightarrow y_w^{\text{motion}}, y_l \rightarrow y_l^{\text{motion}}, t)$, where y_w^{motion} and y_l^{motion} represent structural information extracted from y_w and y_l , respectively.

5.4 Semantic-level Consistency Preference Optimization

Most prior DPO works in diffusion models primarily rely on visual alignment for optimization. However, we propose leveraging language to describe visual differences more precisely, providing specific referential information at the textual semantic level. Inspired by DPO in LMs [61], we introduce the semantic-level consistency alignment by further modeling the semantic information of videos at the textual level. Specifically, for each preference video pair y_w and y_l , we first consider the prompt x as the preferred caption c_w . A VLM (*i.e.*, Qwen2.5-VL [4]) is then tasked to modify c_w based on y_l , adjusting the inconsistent parts to generate c_l . This process ensures that c_w and c_l only differ in textual descriptions where the videos are inconsistent, thereby facilitating more targeted optimization. The $u_{\text{Semantic}}(c_w, c_l, y_w, t)$ within $\mathcal{L}_{\text{Semantic}}$ is then formulated as:

$$u_{\text{Semantic}} = \log \frac{p_{\theta}(y_w | c_w, t)}{p_{\text{ref}}(y_w | c_w, t)} - \log \frac{p_{\theta}(y_w | c_l, t)}{p_{\text{ref}}(y_w | c_l, t)}. \quad (8)$$

By integrating instance, state, motion, and semantic-level preference optimization, the overall loss function for PhysHPO is defined as follows:

$$\mathcal{L}_{\text{PhysHPO}} = \mathcal{L}_{\text{Instance}} + \lambda \mathcal{L}_{\text{State}} + \rho \mathcal{L}_{\text{Motion}} + \mu \mathcal{L}_{\text{Semantic}}, \quad (9)$$

where λ , ρ , and μ denote the loss weights.

6 Experiments

In this section, we conduct extensive experiments to answer the following research questions: **(RQ1)** Does PhysHPO enhance the physical plausibility of generated videos? **(RQ2)** Does PhysHPO compromise other performance aspects? **(RQ3)** How sensitive is PhysHPO to its key components? **(RQ4)** Is PhysHPO more advantageous than SFT for efficiency, effectiveness, and generalizability?

6.1 Experimental Settings

Baselines We apply PhysHPO to the advanced models, CogVideoX-2B and 5B [90]. Due to the unavailability of open-sourced code and model weights for SFT-based methods, *i.e.*, WISA [75] and SynVideo [98], we focus our comparisons on the following baselines: PhyT2V [CVPR'25] [85], vanilla DPO [CVPR'24] [72], SFT [90], and the respective base models. Implementation and results on HunyuanVideo [39] are shown in **Appendix §A**.

Evaluations To evaluate the effectiveness of PhysHPO, we adopt benchmarks focusing on two key aspects: **① Physics-focused**: (i) VideoPhy [6] for interactions of solid-solid, solid-fluid, and fluid-fluid. (ii) PhyGenBench [54] for mechanics, optics, thermal, and materials. **② General Capability**: VBench [34] for both overall quality and semantic.

Implementation Details We train base models on our selected dataset with a global batch size of 8, using the AdamW optimizer and a learning rate of $2e - 5$. Instance-level non-preferred weights are set to $\beta_{\text{err}} = 0.7$ and $\beta_{\text{gap}} = 0.3$, with $N = 2$ for state-level samples. The loss weights λ , ρ , and μ are set to 0.4, 0.3, and 0.2, respectively. All experiments are conducted on 8 NVIDIA H100 GPUs.

6.2 Physical & General Performance of PhysHPO

To answer **RQ1** and **RQ2**, we comprehensively compare PhysHPO with three baseline methods on physics-focused and general benchmarks (Table 2), along with the qualitative results and user study shown in Figure 1, 4 and Figure 5 (Left), respectively. Our observations are summarized as follows:



Figure 4: Qualitative results of PhysHPO. Due to space limitations, results of other baselines are provided in **Appendix §D**. Text prompts are sourced from PhyGenBench [54]. *(1st) Materials*: A delicate, fragile egg is hurled with significant force towards a rugged, solid rock surface, where it collides upon impact. *(2nd) Mechanics*: A vibrant, elastic tennis ball is thrown forcefully towards the ground, capturing its dynamic interaction with the surface upon impact. *(3rd) Optics*: A magnifying glass is gradually moving closer to a coin, revealing the intricate details and textures of the embossed design as it approaches. *(4th) Thermal*: A timelapse captures the transformation of milk in a kettle as the temperature rapidly rises above 100°C.

Obs.④ PhysHPO demonstrates superior performance in enhancing both physical fidelity and general quality. As illustrated in Table 2, our PhysHPO consistently outperforms baseline methods (i.e., PhyT2V [85] and vanilla DPO [72]), which exhibit minimal or even negative performance gains, across three physics-focus benchmarks, each targeting distinct dimensions of physical plausibility. Table 2 further highlights its robustness on VBench [34] for general capabilities. Qualitative comparisons in Figure 1 and Figure 4 provide visual confirmation of PhysHPO’s ability, showcasing clear advantages over the base model. **Obs.⑤ PhysHPO effectively align video generation with human preference.** Figure 5 (Left) shows the user study conducted on both physical and general dimensions, where PhysHPO consistently outperforms baselines in aligning video generation with human preferences, demonstrating its superiority in preference alignment.

6.3 Ablation Analysis

To answer **RQ3**, we conduct evaluations on VideoPhy [6] for the contributions of PhysHPO’s each level and their combinations, as shown in Table 3. We give the following observations: **Obs.⑥ Effectiveness of Hierarchical Preference Optimization.** Table 3 demonstrates a steady performance decline as $\mathcal{L}_{\text{Semantic}}$, $\mathcal{L}_{\text{Motion}}$, and $\mathcal{L}_{\text{State}}$ are progressively removed. This underscores the effectiveness

Table 3: Ablation study on level losses.

Method	SS↑	SF↑	FF↑	Over↑
PhysHPO	32.4	54.1	58.9	45.9
w/o $\mathcal{L}_{\text{Semantic}}$	31.7	53.4	57.1	45.1
w/o $\mathcal{L}_{\text{Semantic}}, \mathcal{L}_{\text{Motion}}$	30.3	52.1	55.4	43.6
Only w/ $\mathcal{L}_{\text{Instance}}$	28.9	50.7	51.8	41.9
Vanilla DPO w/ Our Data	28.2	50.0	51.8	41.3

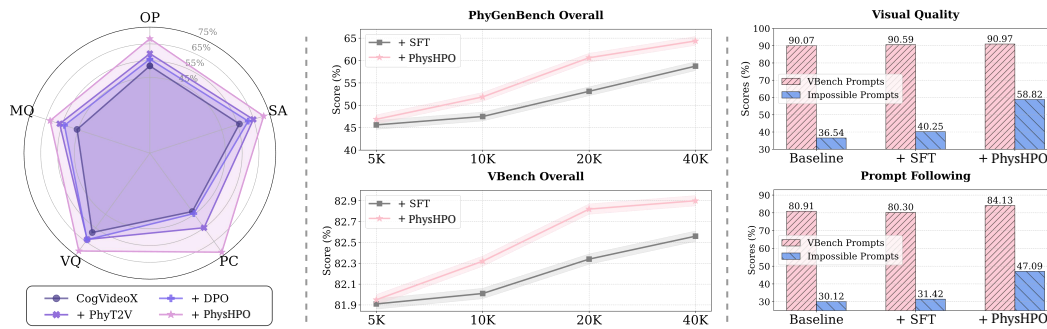


Figure 5: **(Left)** User study across five dimensions: overall preference, semantic adherence, physical commonsense, visual quality, and motion quality. **(Middle)** Performance comparison between PhysHPO and SFT under varying data volumes. **(Right)** Robustness testing with IPV-TXT [5].

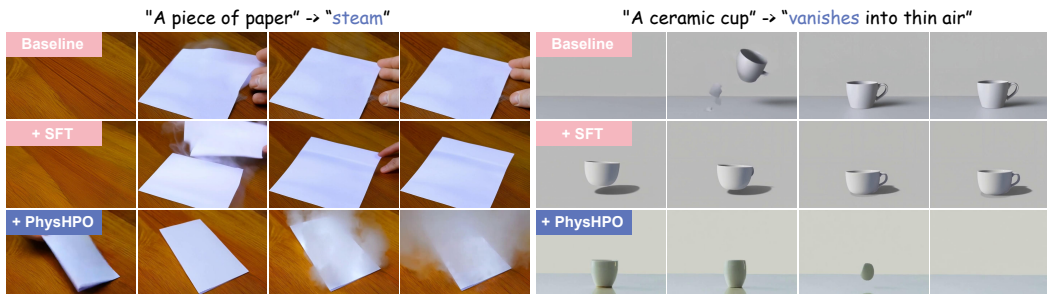


Figure 6: Robustness testing demonstration. Detailed prompts can be found in **Appendix §C**.

of the hierarchical preference optimization framework, where each level contributes uniquely to improving physical fidelity. **Obs. 7 Importance of Fine-grained Alignment.** Fine-grained alignment, both explicit and implicit, is critical for optimization: *i*) The removal of instance-level gap video alignment ("Only w/ $\mathcal{L}_{\text{Instance}}$ " vs. "Vanilla DPO w/ Our Data") highlights the importance of visually explicit modeling in bridging gaps between generated and real-world semantics and dynamics. *ii*) The exclusion of $\mathcal{L}_{\text{Semantic}}$ reveals the necessity of implicit text-video alignment for capturing nuanced and coherent relationships between textual prompts and generated videos.

6.4 Analysis of PhysHPO vs. SFT

To answer **RQ4**, we compare PhysHPO and SFT using the same data on PhyGenBench for physical fidelity and VBench for general quality, as shown in Figure 5 (Middle). Inspired by [5], we further assess both methods with impossible prompts (e.g., "A car drives through the ocean as if it were flying") to evaluate whether improvements in physical fidelity also enable better handling of imaginative or physically impossible scenarios, with results presented in Figure 5 (Right) and Figure 6. We give the following observations: **Obs. 8 PhysHPO is a data-efficient video generation enhancer.** Figure 5 (Middle) demonstrates that PhysHPO consistently outperforms SFT across varying data volumes in both physics-focused and general quality metrics. This reflects PhysHPO's ability to achieve superior performance by leveraging training data more efficiently through fine-grained information utilization. **Obs. 9 PhysHPO enables creative generalization beyond physical laws.** As shown in Figure 5 (Right) and Figure 6, PhysHPO performs better on impossible prompts, generating videos that align more closely with semantic intent while maintaining internal consistency. This suggests that enhancing physical fidelity not only improves adherence to real-world physics but also equips the model with a stronger ability to generalize and adapt to imaginative scenarios, demonstrating learning beyond simple physical rules.

7 Conclusion

In this paper, we propose **PhysHPO**, a novel framework for Hierarchical Cross-Modal Direct Preference Optimization, enhancing video fine-grained alignment across four levels: instance, state, motion, and semantic. Recognizing real-world videos as the best reflections of physical phenomena, we introduce an automated data selection pipeline to efficiently utilize large-scale text-video datasets, removing the need for exhaustive dataset construction. Extensive evaluations on both physics-focused and general benchmarks demonstrate that PhysHPO significantly improves the physical plausibility and overall quality of existing video generation models, addressing key challenges in physically plausible video generation.

Acknowledgement

This research was supported by the Innovation and Technology Fund of HKSAR under grant number GHX/054/21GD.

References

- [1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semd-edup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.
- [2] 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.
- [3] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-v1 technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [5] Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. *arXiv preprint arXiv:2503.14378*, 2025.
- [6] 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.
- [7] David Brandfonbrener, Hanlin Zhang, Andreas Kirsch, Jonathan Richard Schwarz, and Sham Kakade. Color-filter: Conditional loss reduction filtering for targeted language model pre-training. *Advances in Neural Information Processing Systems*, 37:97618–97649, 2024.
- [8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>.
- [9] Miaomiao Cai, Simiao Li, Wei Li, Xudong Huang, Hanting Chen, Jie Hu, and Yunhe Wang. Dspo: Direct semantic preference optimization for real-world image super-resolution. *arXiv preprint arXiv:2504.15176*, 2025.
- [10] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.
- [11] Haodong Chen, Lan Wang, Harry Yang, and Ser-Nam Lim. Omnicreator: Self-supervised unified generation with universal editing. *arXiv preprint arXiv:2412.02114*, 2024.
- [12] Harold Haodong Chen, Harry Yang, and Ser-Nam Lim. Beyond generation: Unlocking universal editing via self-supervised fine-tuning. *arXiv preprint arXiv:2412.02114*, 2024.
- [13] Harold Haodong Chen, Haojian Huang, Xianfeng Wu, Yexin Liu, Yajing Bai, Wen-Jie Shu, Harry Yang, and Ser-Nam Lim. Temporal regularization makes your video generator stronger. *arXiv preprint arXiv:2503.15417*, 2025.
- [14] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

- [15] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, Ting-Che Lin, Shilong Zhang, Fu Li, Chuan Li, Xing Wang, Yanghua Peng, Peize Sun, Ping Luo, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025.
- [16] Haoran Cheng, Qide Dong, Liang Peng, Zhizhou Sha, Weiguo Feng, Jinghui Xie, Zhao Song, Shilei Wen, Xiaofei He, and Boxi Wu. Discriminator-free direct preference optimization for video diffusion. *arXiv preprint arXiv:2504.08542*, 2025.
- [17] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023.
- [18] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.
- [19] Karan Dalal, Daniel Kocaja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, et al. One-minute video generation with test-time training. *arXiv preprint arXiv:2504.05298*, 2025.
- [20] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.
- [21] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What’s in my big data? *arXiv preprint arXiv:2310.20707*, 2023.
- [22] Yutang Feng, Sicheng Gao, Yuxiang Bao, Xiaodi Wang, Shumin Han, Juan Zhang, Baochang Zhang, and Angela Yao. Wave: Warping ddim inversion features for zero-shot text-to-video editing. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.
- [23] Jinlan Fu, huangfushenzhen, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. CHip: Cross-modal hierarchical direct preference optimization for multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=71pDn2MhM2>.
- [24] Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Hongxia Ma, Li Zhang, Boxing Chen, Hao Yang, et al. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. *arXiv preprint arXiv:2402.18191*, 2024.
- [25] David Grangier, Simin Fan, Skyler Seto, and Pierre Abblin. Task-adaptive pretrained language models via clustered-importance sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=p6ncr0eTKE>.
- [26] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- [27] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.
- [28] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025.
- [29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

- [30] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.
- [31] Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *arXiv preprint arXiv:2504.13122*, 2025.
- [32] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- [33] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025.
- [34] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [35] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.
- [36] Lifan Jiang, Boxi Wu, Jiahui Zhang, Xiaotong Guan, and Shuang Chen. Huidpo: Enhancing video generation through direct preference optimization for human-centric alignment. *arXiv preprint arXiv:2502.01690*, 2025.
- [37] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.
- [38] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.
- [39] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [40] Simon Le Cleac’h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8(5):2780–2787, 2023.
- [41] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025.
- [42] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.
- [43] Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. *arXiv preprint arXiv:2503.12689*, 2025.
- [44] Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis. *arXiv preprint arXiv:2504.13129*, 2025.

- [45] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*, 2023.
- [46] Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16189–16211, 2024.
- [47] Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *arXiv preprint arXiv:2402.00530*, 2024.
- [48] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.
- [49] Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Simon Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, Philip S. Yu, and Meng Cao. TIS-DPO: Token-level importance sampling for direct preference optimization with estimated weights. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=oF6e2WwxX0>.
- [50] Runtao Liu, Chen I Chieh, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*, 2024.
- [51] Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024.
- [52] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024.
- [53] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=BTKAeLqLMw>.
- [54] 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.
- [55] Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.
- [56] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025.
- [57] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.
- [58] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [59] Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. *arXiv preprint arXiv:2405.12915*, 2024.

- [60] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.
- [61] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- [62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [63] Shouwei Ruan, Zhenyu Wu, Yao Huang, Ruochen Zhang, Yitong Sun, Caixin Kang, and Xingxing Wei. Towards nsfw-free text-to-image generation via safety-constraint direct preference optimization. *arXiv preprint arXiv:2504.14290*, 2025.
- [64] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.
- [65] Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, and Binglu Wang. Finephys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1905–1916, 2025.
- [66] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.
- [67] Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*, 2024.
- [68] Xin Tan, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, and Hong Xu. Dsv: Exploiting dynamic sparsity to accelerate large-scale video dit training. *arXiv preprint arXiv:2502.07590*, 2025.
- [69] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.
- [70] Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36:53983–53995, 2023.
- [71] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [72] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.
- [73] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- [74] Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*, 2024.
- [75] 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.

- [76] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.
- [77] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [78] Xianfeng Wu, Yajing Bai, Haoze Zheng, Harold Haodong Chen, Yexin Liu, Zihao Wang, Xuran Ma, Wen-Jie Shu, Xianzu Wu, Harry Yang, et al. Lightgen: Efficient image generation through knowledge distillation and direct preference optimization. *arXiv preprint arXiv:2503.08619*, 2025.
- [79] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.
- [80] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024.
- [81] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=uPSQv01eAu>.
- [82] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Toonrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024.
- [83] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024.
- [84] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.
- [85] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. *arXiv preprint arXiv:2412.00596*, 2024.
- [86] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [87] Nianzu Yang, Pandeng Li, Liming Zhao, Yang Li, Chen-Wei Xie, Yehui Tang, Xudong Lu, Zhihang Liu, Yun Zheng, Yu Liu, et al. Rethinking video tokenization: A conditioned diffusion-based approach. *arXiv preprint arXiv:2503.03708*, 2025.
- [88] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlippi: Towards physically plausible video generation with vision and language informed physical prior. *arXiv e-prints*, pages arXiv–2503, 2025.
- [89] Yu Yang, Aaditya K Singh, Mostafa Elhoushi, Anas Mahmoud, Kushal Tirumala, Fabian Gloeckle, Baptiste Rozière, Carole-Jean Wu, Ari S Morcos, and Newsha Ardalani. Decoding data quality via synthetic corruptions: Embedding-guided pruning of code data. *arXiv preprint arXiv:2312.02418*, 2023.

- [90] 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.
- [91] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.
- [92] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025.
- [93] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [94] Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159*, 2024.
- [95] Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. Tagcos: Task-agnostic gradient clustered coreset selection for instruction tuning data. *arXiv preprint arXiv:2407.15235*, 2024.
- [96] Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.
- [97] Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhenghong Liu, and Hao Zhang. Fast video generation with sliding tile attention, 2025. URL <https://arxiv.org/abs/2502.04507>.
- [98] Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyang Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis. *arXiv preprint arXiv:2503.20822*, 2025.
- [99] 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.
- [100] Xiaojing Zhong, Xinyi Huang, Xiaofeng Yang, Guosheng Lin, and Qingyao Wu. Deco: Decoupled human-centered diffusion video editing with motion consistency. In *European Conference on Computer Vision*, pages 352–370. Springer, 2024.
- [101] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.
- [102] Bingwen Zhu, Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Yidi Wu, Huyang Sun, and Zuxuan Wu. Aligning anime video generation with human feedback. *arXiv preprint arXiv:2504.10044*, 2025.
- [103] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Both the abstract and introduction effectively highlight the paper's pioneering exploration of fine-grained direct preference optimization modeling and data selection techniques in the field of video generation. We also claim the contributions validated through experimental results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper includes a discussion of its limitations in **Appendix §E**, where we outline the constraints of current work and propose potential directions for future research.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not involve theoretical results. However, we clearly outline the relevant formulas in the Methodology section.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed information on the hyperparameter settings and related analyses necessary for reproducing the main experimental results. Additionally, the code is made available in an anonymous manner.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The complete code, based on the primary model CogVideoX, is provided anonymously. The data used is from the open-source dataset OpenVid.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides detailed information on data selection, hyperparameter settings, and the type of optimizer used, ensuring clarity in understanding the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Following previous works, the paper does not report error bars or provide information about the statistical significance of the experiments. Besides, due to the long training times and high computational costs associated with their experiments, we do not conduct multiple runs to calculate error bars or perform statistical significance tests.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides sufficient information on the computer resources needed to reproduce the experiments. The authors mention that all training for PhysHPO was conducted on eight NVIDIA H100 GPUs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This paper discuss the Broader Impacts in the Conclusion Section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: The authors discuss Ethical Implications in **Appendix §F**.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: The paper properly credits the creators of the datasets and models used, and it mentions the licenses and terms of use, ensuring compliance with ethical guidelines and legal requirements.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: The new assets introduced in the paper are well documented, and the documentation is provided alongside the assets, ensuring clarity and reproducibility.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: The User Study was conducted by recruiting participants within the institution rather than through a crowdsourcing platform. However, this provides full text of instructions given to participants of the user study in **Appendix §C**.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs in terms of training or fine-tuning as part of its core contributions. The use of LLMs is limited to the data selection process and for tasks such as editing the manuscript. These uses do not impact the originality or core methodology of the research, and therefore do not require detailed declaration.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A More Results on HunyuanVideo

To further validate the superiority of our proposed PhysHPO, in this section, we further apply PhysHPO on HunyuanVideo-540p [39] implemented by FastVideo [97].

Method	VideoPhy [6]				PhyGenBench [54]					VBench [34]		
	SS↑	SF↑	FF↑	Over.↑	Mechanics↑	Optics↑	Thermal↑	Materials↑	Overall↑	Total↑	Quality↑	Semantic↑
HunyuanVideo [39]	19.7	43.2	42.9	33.4	37.5	58.0	36.7	45.0	45.6	81.4	83.1	74.9
+ PhyT2V [85]	21.1	45.9	50.0	36.3	40.0	58.0	33.3	47.5	46.3	80.5	82.0	74.5
+ PhysHPO (Ours)	26.8	50.7	55.4	41.6	47.5	62.0	43.3	60.0	54.4	81.9	83.5	75.7

Table 4: Evaluation of PhysHPO on HunyuanVideo [39].

Quantitative Results Table 4 presents the quantitative results of PhysHPO on HunyuanVideo. Consistent with our **Obs. 4** in the main content, PhysHPO achieves superior performance in enhancing both physical fidelity and overall video quality. These results further validate its effectiveness.

Qualitative Results We demonstrate results in Figure 7. More results on *human action* are shown in Section §D.

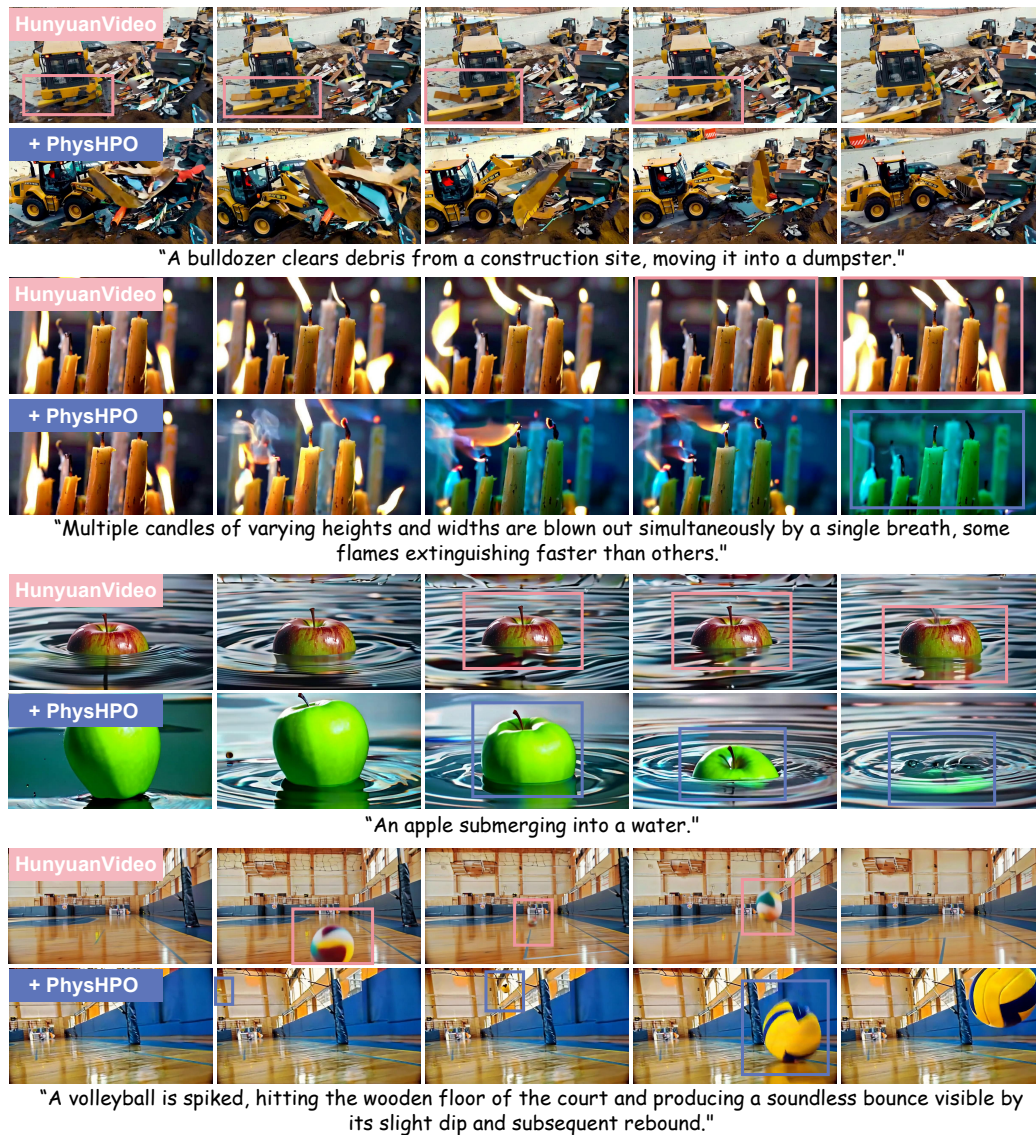


Figure 7: Qualitative demonstration of PhysHPO on HunyuanVideo [39].

B More Details of Data Selection

In-depth Evolving Prompting To achieve more fine-grained scoring for the physical fidelity of real-world videos, we employ the in-depth evolving prompting strategy [84]. This approach augments sample captions to emphasize specific physical phenomena. An example of our crafted *evolving-dynamic* prompt is shown in Figure 8.

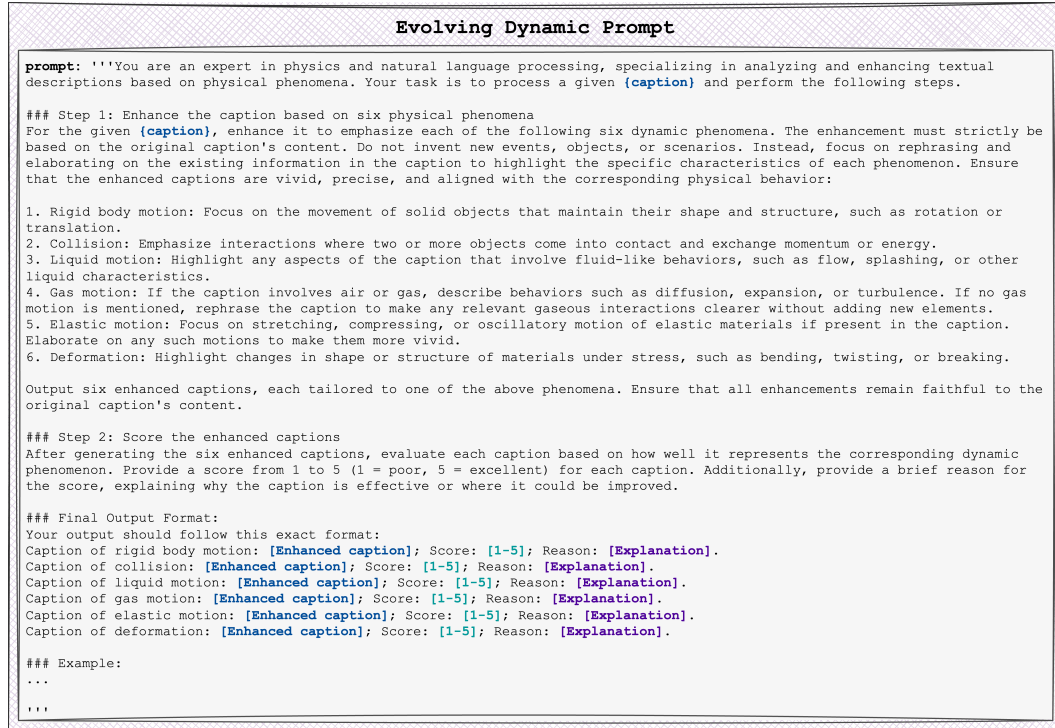


Figure 8: Demonstration of crafted *evolving-dynamic* prompt.

Similarly, the *evolving-thermodynamic* and *evolving-optic* prompts follow the same structure.

In addition to the performance comparison between our evolving-based scoring and direct scoring shown in Figure 3, we also provide examples with detailed scores under different strategies in Figure 9.

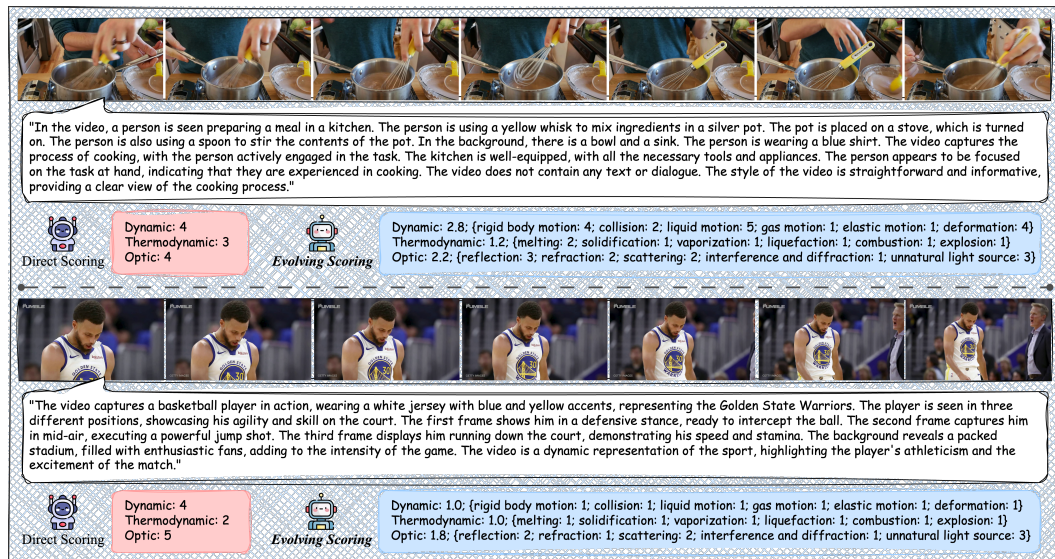


Figure 9: Comparison of direct scoring vs. our evolving scoring.

Specifically, we compare a sample of *dynamic* stirring (*Top*) with another sample where the camera zooms out but the scene remains *static* (*Bottom*). As shown, the direct scoring strategy assigns similar scores to these two distinct videos, while our evolving-based scoring provides more fine-grained differentiation and accurate assessments.

Analysis of Caption-based Selection In our data selection process, we first performed *video-based* filtering to ensure reality, followed by *caption-based* filtering to enhance physical fidelity and diversity. Here, we further validate the efficiency and effectiveness of our caption-based selection strategy.

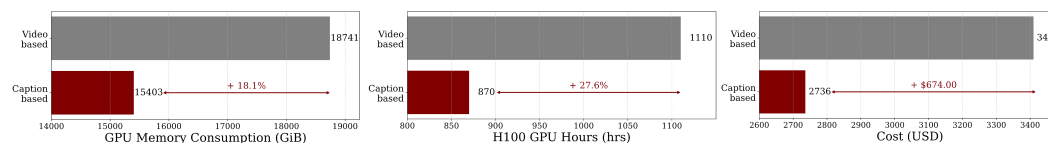


Figure 10: Efficiency comparison of video-based selection vs. our caption-based selection. (**Left**) GPU Memory Consumption (GiB). (**Middle**) H100 GPU Hours (hrs). (**Right**) Cost (USD).

Figure 10 first demonstrates an efficiency comparison of our caption-based selection and video-based selection. Specifically, ❶ GPU Memory Consumption (*Left*): Caption-based selection requires significantly less GPU memory, reducing usage by 18.1% compared to video-based selection (15,403 GiB vs. 18,741 GiB). This reduction highlights the computational efficiency of our method in terms of memory footprint. ❷ H100 GPU Hours (*Middle*): Video-based selection increases compute time by 27.6% (3,410 hrs vs. 2,736 hrs), indicating that caption-based selection is more time-efficient and better suited for large-scale processing. ❸ Cost (*Right*): At scale, video-based workflows incur an additional \$674.00 in cost compared to caption-based workflows, further emphasizing the cost-efficiency of our approach.

To further validate the effectiveness of our caption-based selection, we randomly sampled 100 data points from the reality-selected data pool for human scoring. We then analyze the scoring accuracy by comparing human scoring results with three scoring methods: video-based scoring, caption-based direct scoring, and our caption-based evolving scoring. As demonstrated in Figure 11, our caption-based evolving scoring achieves the highest alignment with human scoring across all three evaluated dimensions. These results further confirm that our method not only improves computational efficiency but also ensures superior scoring reliability.

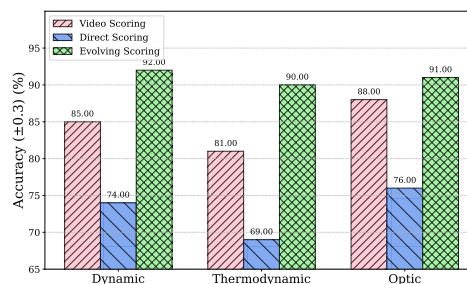


Figure 11: Adjacency accuracy (± 0.3) of different scoring strategies with human scoring.

Statistical Overview To provide a clearer understanding of each stage in our data selection process, we present the statistical breakdown in Table 5. The raw data is sourced directly from OpenVidHD-0.4M [57]. Moving forward, we will process more existing high-quality text-video datasets (e.g., MiraData [37], InternVid [76]) using our data selection pipeline to facilitate future research.

Table 5: Statistics of data selection.

Data Selection	Total Count
Raw Data	433,523
+ Reality	266,931
+ Physics	59,076
+ Diversity	21,085

Data Sample Demonstration Real-world videos serve as direct reflections of physical phenomena. Here, we present examples of our selected videos as illustrative examples, as shown in Figure 12. These examples highlight diverse physical scenarios, providing a reference for the types of phenomena captured in our dataset.

C More Experimental Settings and Analysis

C.1 More Details of Experimental Settings

Physics-focused Metric We evaluate the physical fidelity of generated videos using two widely adopted physics-focused benchmarks: VideoPhy [ICLR'25] [6] and PhyGenBench [ICML'25] [54].

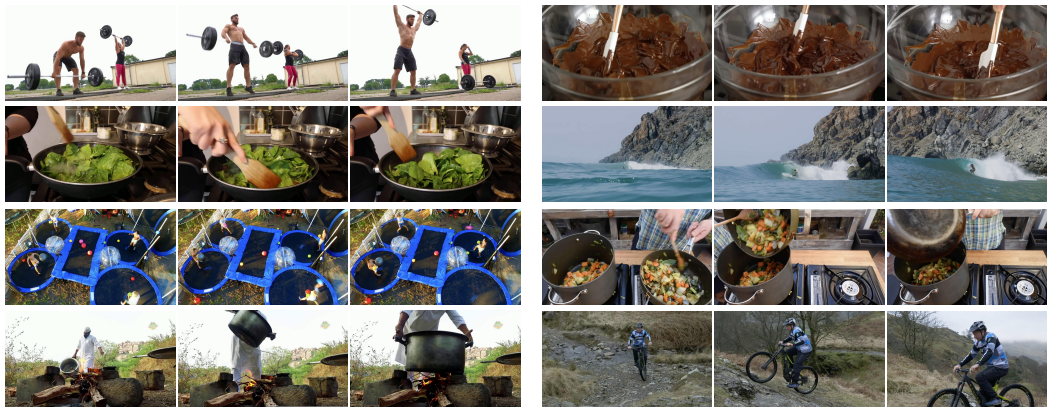


Figure 12: Randomly sampled data demonstration from our selected dataset.

Following the evaluation protocol introduced in WISA [75], we leverage VideoCon-Physics from VideoPhy [6] to assess two key metrics: **Physical Law Consistency (PC)** and **Semantic Coherence (SA)** of the generated videos. Specifically, we use 344 carefully curated prompts from VideoPhy and 160 prompts from PhyGenBench, both designed to reflect a diverse range of physical principles and scenarios.

To quantify performance, we follow the scoring criteria outlined in the respective benchmark papers. For both VideoPhy and PhyGenBench, we consider PC and SA values greater than or equal to 0.5 as $PC = 1$ and $SA = 1$, while values less than 0.5 are treated as $PC = 0$ and $SA = 0$. Specifically:

- For **VideoPhy**, we report the proportion of generated videos where both $PC = 1$ and $SA = 1$, reflecting the alignment with physical laws and semantic intent simultaneously.
- For **PhyGenBench**, we focus exclusively on the Physical Commonsense Alignment (*PCA*) score, which evaluates the consistency of physical reasoning under the condition that $PC = 1$.

This evaluation framework ensures consistency with prior work and provides a robust assessment of the physical fidelity and semantic coherence of the generated videos.

Impossible Robustness Testing Inspired by Impossible Videos [ICML'25] [5], we further employ prompts from IPV-TXT [5] to evaluate whether PhysHPO goes beyond merely fitting fixed physical phenomena and demonstrates stronger generalization and robustness. The results are illustrated in Figure 5 (Right) and Figure 6.

For the quantitative comparison shown in Figure 5 (Right), we follow the evaluation protocol in [5] to assess the generated videos on two key metrics:

- **Visual Quality (VQ):** This metric is derived by combining six factors from VBench [34], including Subject Consistency, Background Consistency, Motion Smoothness, Aesthetic Quality, Imaging Quality, and Dynamic Degree. These factors are aggregated into a single score to reflect the overall visual quality of the generated videos.
- **Impossible Prompt Following (IPF):** This metric evaluates how well the generated videos align with the semantic intent of the impossible prompts. Following [5], we utilize GPT-4o to provide a binary judgment for each video based on prompt adherence and calculate the proportion of positive judgments as the final IPF score.

For the qualitative comparison presented in Figure 6, we further detail the prompts here:

- **Left:** "A piece of paper mysteriously transforms into steam on a wooden table. The surreal sequence begins with someone placing the paper down, followed by a gentle touch from a human hand that triggers an unexpected reaction - the solid paper instantly vaporizes into wisps of white steam that dissipate into the air."
- **Right:** "A ceramic cup sitting on a plain table surface mysteriously vanishes into thin air in this photo-realistic footage. The abrupt disappearance happens without any visible cause, defying physics in an uncanny way. The scene is captured in crisp, lifelike detail with natural lighting."

User Study We conduct a user study to evaluate human preferences using both the mean opinion score (MOS) and direct pairwise comparisons. Specifically, we design a user-friendly interface to

facilitate the evaluation process and collect feedback from a total of 15 volunteers. The detailed instructions provided to participants are shown below.

User Study: Physically Plausible Video Generation

Thank you for participating in our user study! Please follow these steps to complete your evaluation:

1. **Video Generation:** Carefully read the target prompt provided, and then view the provided videos.
2. **Scoring Criteria:** Assign a score to each generated video based on the following aspects (1 being the lowest, 5 being the highest):
 - **Overall Preference:** A holistic evaluation of your overall satisfaction with the generated video, including aspects such as semantics, physical reasoning, visual quality, and motion quality.
 - **Semantic Adherence:** How accurately the video reflects the input semantic description or textual instructions.
 - **Physical Commonsense:** Whether the video content aligns with basic physical commonsense, such as the laws of motion, object interactions, and logical behavior.
 - **Visual Quality:** The visual appearance of the video, including resolution, clarity, color representation, and texture details.
 - **Motion Quality:** The smoothness and naturalness of motion in the video, including object or character trajectories and speed variations.
3. **Submission:** Click the "Submit Scores" button to submit your scores.

Notations:

1. We observe that the edge browser is not fully compatible with our interface. Chrome is recommended.
2. Remember to click the "Submit Scores" button after your evaluation.
3. If you see that videos and the score sliders are not aligned, shrinking your page usually works.
4. If the video seems to be stuck, usually waiting for a few seconds will solve this.
5. If the page is not responsive for a long time, please try to refresh it.
6. If you have any questions, please directly contact us. Thank you for your time and effort!

C.2 More Analysis

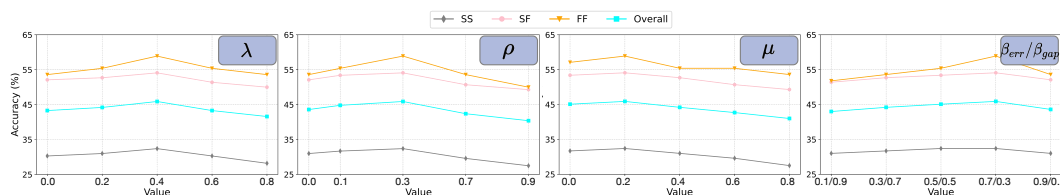


Figure 13: Hyperparameter analysis of PhysHPO on VideoPhy [6].

Analysis of Hyperparameters Figure 13 presents a detailed analysis of the impact of various hyperparameters on the performance of PhysHPO on the VideoPhy [6] benchmark. *a)* The first plot examines the state-level loss weight (λ), showing that accuracy improves as λ increases, peaking at $\lambda = 0.4$, before declining. This suggests that moderate emphasis on state-level consistency enhances performance, while excessive weighting may lead to overfitting to low-level details. *b)* Similarly, the motion-level loss weight (ρ) in the second plot demonstrates a comparable trend, with performance reaching its best at $\rho = 0.3$. This highlights the importance of capturing motion dynamics, but also the potential detriment of overemphasizing this aspect. *c)* The third plot explores the semantic-level loss weight (μ), where accuracy improves with increasing μ up to $\mu = 0.2$, after which the performance plateaus or slightly declines. This underscores the necessity of aligning semantic-level information to ensure coherent and contextually accurate video generation. *d)* Finally, the fourth plot analyzes the balance between error (β_{err}) and gap (β_{gap}) samples in instance-level negative sampling. Results

indicate that the weighting ($\beta_{\text{err}}/\beta_{\text{gap}} = 0.7/0.3$) achieves the highest accuracy, as both sample types contribute to model robustness. Overemphasizing either error ($\beta_{\text{err}}/\beta_{\text{gap}} = 0.9/0.1$) or gap samples ($0.1/0.9$) diminishes performance, likely due to an imbalanced training signal.

Analysis of Non-preferred Samples The choice of non-preferred samples determines their contrast with preferred samples and significantly impacts the model’s focus on preference learning. Here, we further explore the selection strategies for non-preferred samples at the instance, state, and motion levels, while ensuring consistency with the corresponding preferred sample configurations in the main text. Specifically, *i*) **Instance-level**: The number of generated error videos corresponding to a preferred video, *i.e.*, how many videos are generated to select the error video; *ii*) **State-level**: The number of boundary frames swapped to construct non-preferred samples; *iii*) **Motion-level**: The choice of structural representations, *e.g.*, depth map, optical flow, and Canny edge.

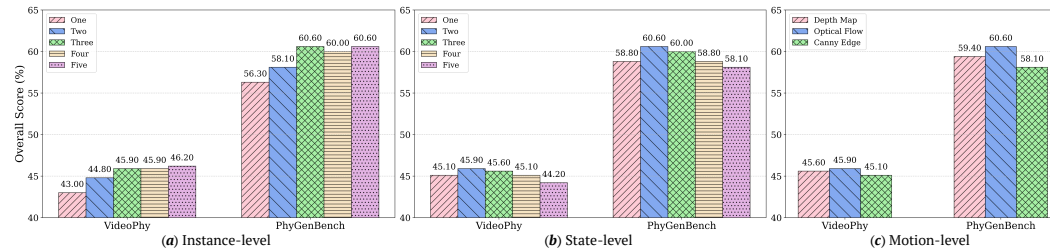


Figure 14: Non-preferred samples analysis of (a) Instance-level, (b) State-level, and (c) Motion-level.

Figure 14 illustrates the impact of different non-preferred sample selection strategies on model performance: (a) Instance-level: We observe that selecting three generated samples achieves the best performance. While increasing the number of samples can lead to slight performance gains, it also incurs higher computational costs, making three a practical balance. (b) State-level: Swapping a moderate number of boundary frames yields optimal results, as excessive or insufficient frame swapping diminishes the contrast between preferred and non-preferred samples. (c) Motion-level: Optical flow achieves the highest overall score, demonstrating its effectiveness in capturing motion dynamics. In contrast, depth maps and Canny edge representations show relatively lower performance, likely due to their less detailed motion information. These results emphasize the importance of carefully balancing computational efficiency and performance when selecting non-preferred samples. Proper configurations at each level ensure effective and robust model performance.

D Exhibition Board

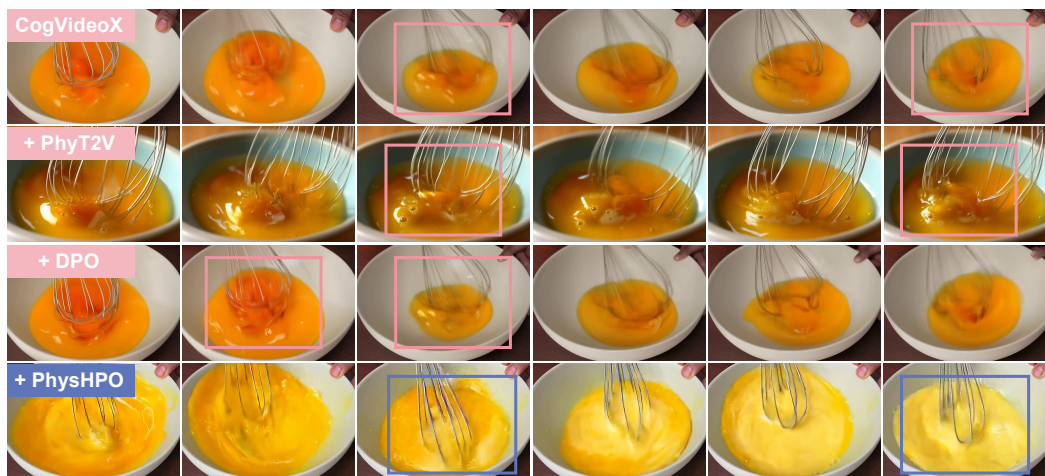
We provide more comparison results here in Figure 15 (with baselines) and Figure 16 (with physically focused prompts), along with results on *human action/motion* in Figure 17 (HunyuanVideo [39]) and Figure 18 (CogVideoX [90]). *We would highly recommend watching the **Webpage**.*

E Limitation and Future Works

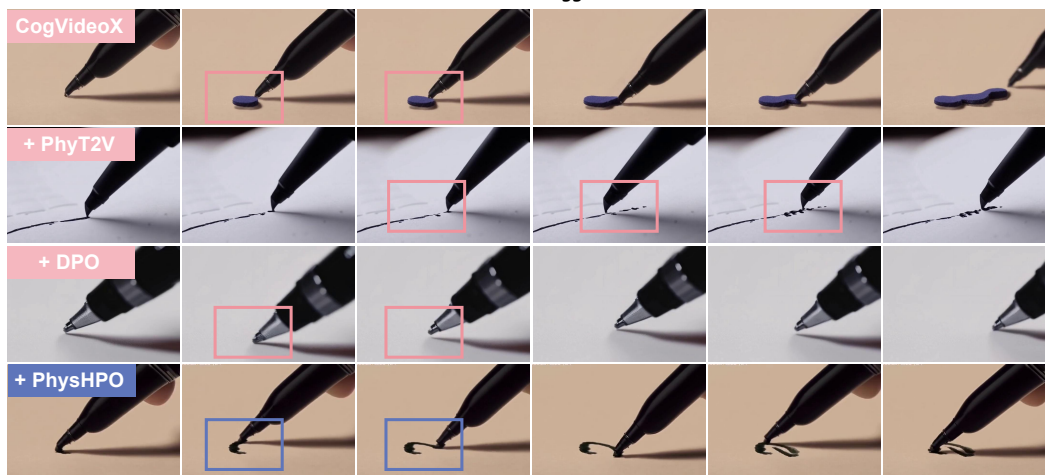
While PhysHPO excels at aligning videos with fine-grained precision, the substantial computational cost required for training large-scale video generation models remains a potential limitation, particularly for individuals and organizations with limited resources. Future research should explore more lightweight architectures (*e.g.*, FramePack [96]) or further explore test-time preference alignment to achieve greater performance improvements with reduced computational overhead. Additionally, the data selection strategy introduced in this paper is designed to be relatively simple for practicality. Future work should further optimize it based on specific task requirements.

F Ethical Implications

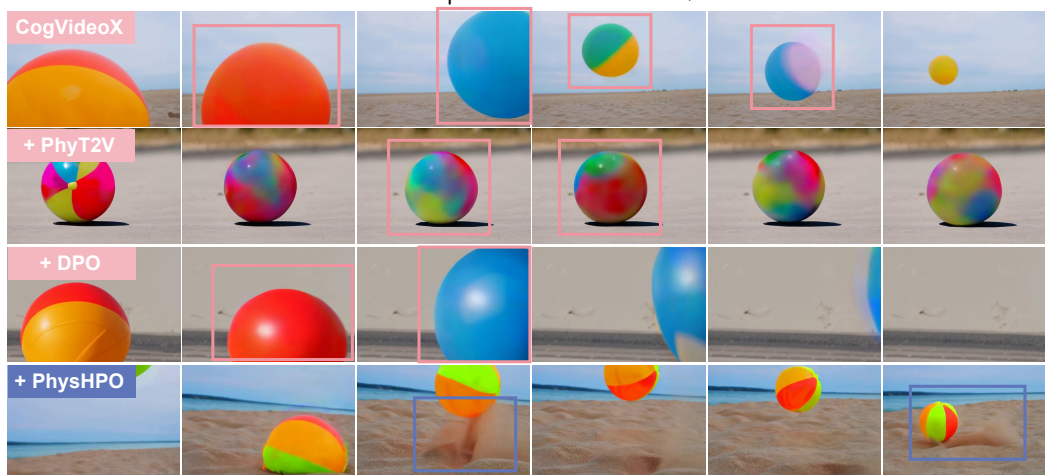
PhysHPO is developed as a hierarchical preference optimization strategy for RESEARCH ONLY. It may still raise important ethical considerations, particularly around content generation. The ability to generate high-quality videos can potentially be misused for creating misleading or harmful content. To mitigate this risk, we recommend incorporating safeguards such as adding watermarks to generated videos to ensure transparency and authenticity. Additionally, guidelines on responsible use should be established, emphasizing its application in ethical and creative contexts, such as educational, artistic, or research-based scenarios, while discouraging its use for deceptive or harmful purposes.



"A whisk mixes an egg in a bowl."



"A black pen is used to write on the smooth, white surface of a notebook, showcasing the interaction between the pen and the notebook surface."



"A vibrant, elastic beach ball is thrown forcefully towards the ground, capturing its dynamic interaction with the surface upon impact."

Figure 15: More comparison demonstrations with baselines.

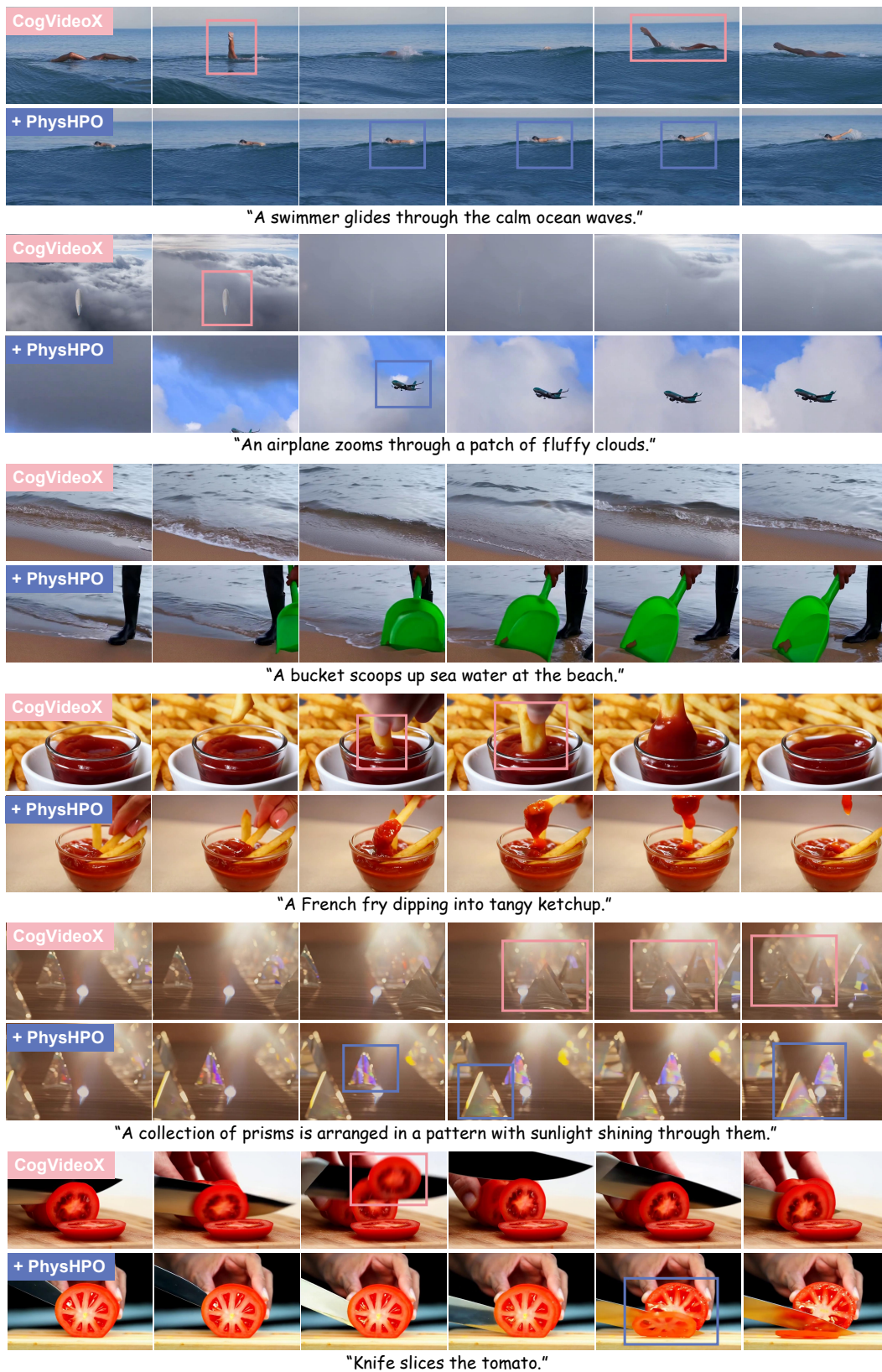


Figure 16: More results demonstrations with prompts sourced from VideoPhy [6].

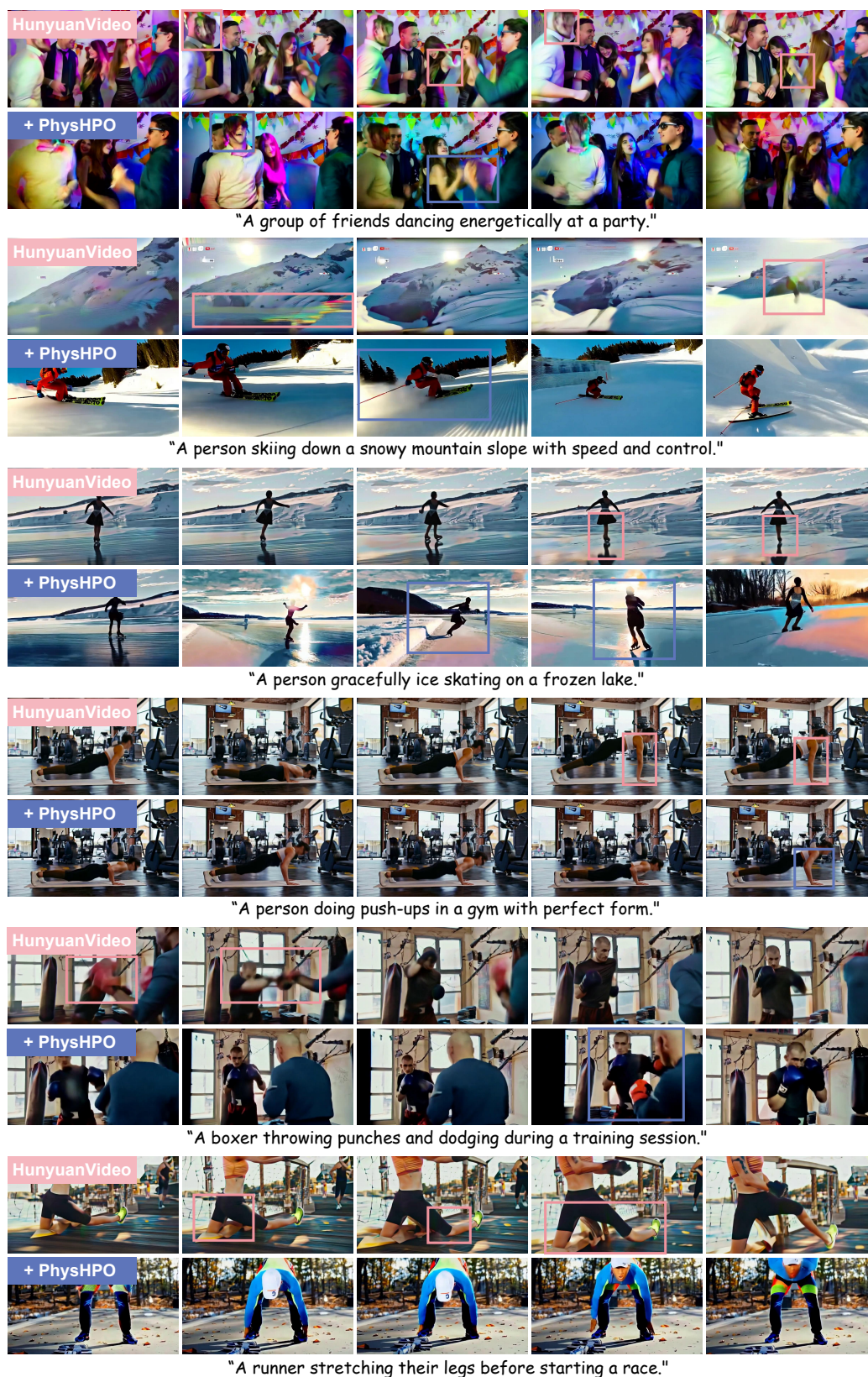


Figure 17: More results demonstrations with human action/motion-focused prompts.

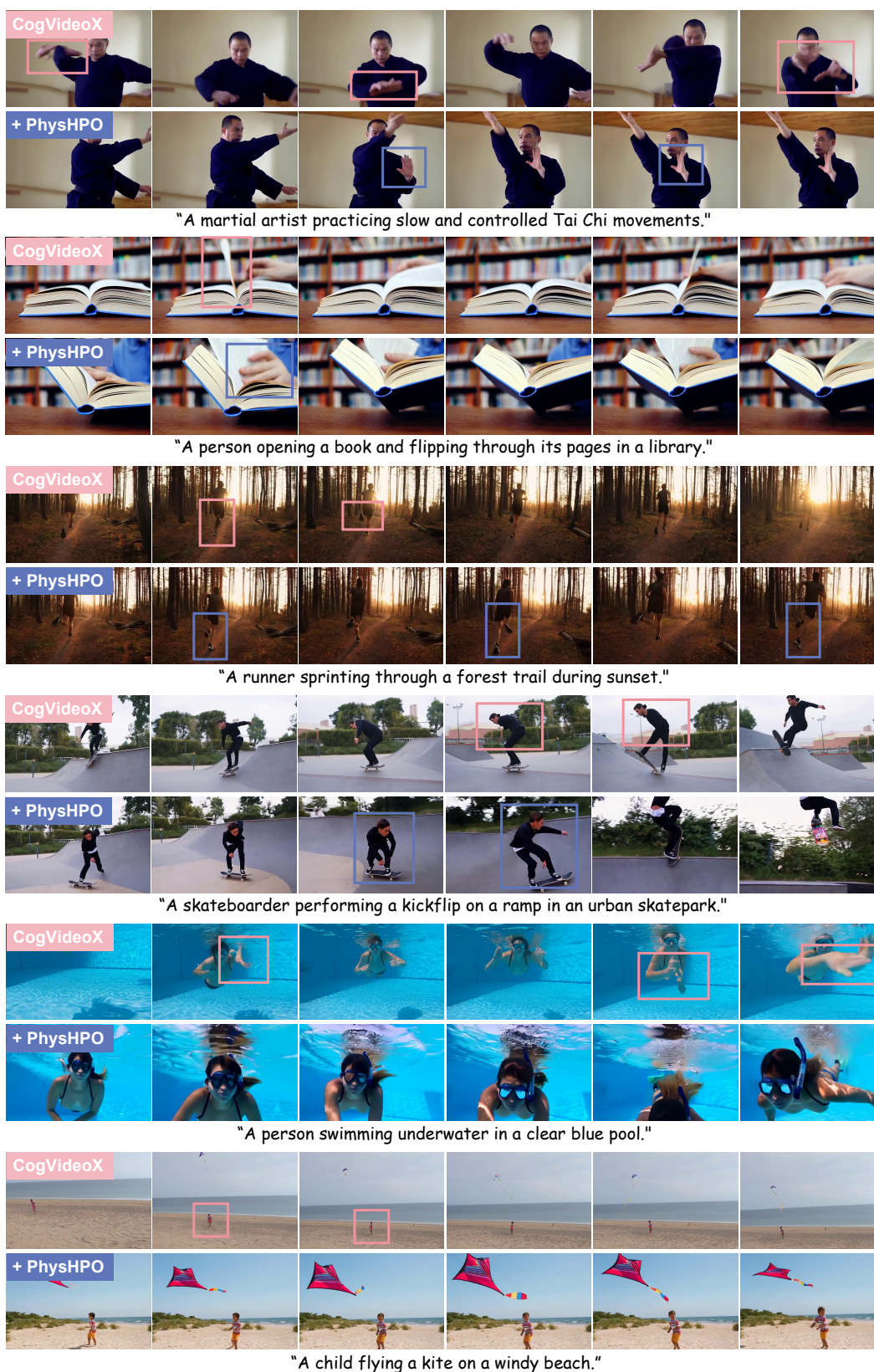


Figure 18: More results demonstrations with human action/motion-focused prompts.