# PhyGDPO: Physics-Aware Groupwise Direct Preference Optimization for Physically Consistent Text-to-Video Generation

Yuanhao Cai[1,2,⋆], Kunpeng Li[1], Menglin Jia[1], Jialiang Wang[1], Junzhe Sun[1], Feng Liang[1], Weifeng Chen[1], Felix Juefei-Xu[1], Chu Wang[1], Ali Thabet[1], Xiaoliang Dai[1], Xuan Ju[3], Alan Yuille[2,†], and Ji Hou[1,4,⋆,†]

[1] Meta, [2] Johns Hopkins University, [3] CUHK, [4] Tencent Hunyuan

**Abstract.** Recent advances in text-to-video (T2V) generation have achie ved good visual quality, yet synthesizing videos that faithfully follow physical laws remains an open challenge. Existing methods mainly based on graphics or prompt extension struggle to generalize beyond simple simulated environments or learn implicit physics reasoning. The scarcity of training data with rich physics interactions and phenomena is also a problem. In this paper, we first introduce a Physics-Augmented video data construction Pipeline, PhyAugPipe, that leverages a vision–language model (VLM) with chain-of-thought reasoning to collect a training dataset, PhyVidGen-135K. Then we formulate a principled Physics-aware Groupwise Direct Preference Optimization, PhyGDPO, framework that uses real-world video as winning case to guarantee correct physics learning and builds upon the groupwise Plackett–Luce probabilistic model to capture holistic preferences beyond pairwise comparisons. We design a Physics-Guided Rewarding (PGR) scheme that uses VLM-based physical rewards to direct the optimization to focus on challenging physics. Plus, we propose a LoRA-Switch Reference (LoRA-SR) scheme that avoids full-model duplication as reference for efficient training. Experiments show that our method outperforms state-of-the-art methods. Our code and data is at https://github.com/caiyuanhao1998/Open-PhyGDPO

**Keywords:** Video Generation · Direct Preference Optimization

## 1 Introduction

With the rapid development of computing power and the increasing scale of training data, text-to-video (T2V) generation models [7,11,65,81] have witnessed significant progress. However, accurately and consistently modeling the physics in the generated video still remains challenging and less explored by existing methods. Improving the physics reasoning capability of video generation models can make them closer to real-world simulators, which can benefit a wide range of applications such as video gaming [10, 62], autonomous driving [69, 74, 79], robotics [2, 3, 16], film making [56, 68, 75], and so on.

⋆ Work was done in Meta Superintelligence Labs, † Corresponding Authors

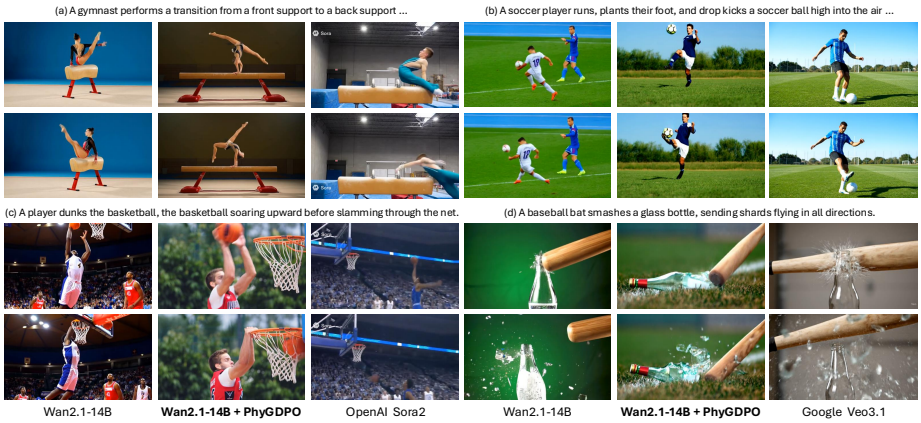


**Fig. 1:** Text-to-video generation on four challenging action categories: (a) gymnastics, (b) soccer, (c) basketball, and (d) glass smashing. When using our post-training method, PhyGDPO, on Wan2.1-T2V-14B [65], the model yields more physically plausible results than two closed-source models: OpenAI Sora2 [7] and Google Veo3.1 [20] by generating well-structured human bodies and reasonable physical interactions such as foot kicking the soccer, basketball passing through the hoop, and glass shattering.

To improve the physics modeling, graphics-based methods [34] rely on simulation engines to specify physical parameters of simple scenarios such as perfectly elastic collisions and basic rigid body dynamics. Yet, they are hard to apply in real-world scenes as the environments are much more complex.

Another technical route [77, 82] is based on prompt extension with a large language model (LLM). These methods use an LLM to extend the prompts with explicit physical laws and phenomena, and then use the extended prompts as input to simulate the physics by iteratively generating the video or finetuning the model on a small subset. These T2V models simply follow the LLM-augmented prompts and lack the ability of thinking in physics. Instead, they use the LLM as their surrogate brain and outsource the reasoning process to it. Even so, the prompt following ability of current T2V models is still limited and existing LLM's physics reasoning is also weak and often erroneous, which may in turn mislead the T2V model when it follows such guidance with bias or errors.

To learn the implicit physics, most T2V foundation models are trained on massive collections of high-quality text–video pairs. However, collecting and annotating such data is extremely costly and labor-intensive. Moreover, models obtained through such supervised fine-tuning (SFT) or training still exhibit limited physics reasoning ability. For instance, OpenAI Sora2 [7] and Google Veo3.1 [20] often fail in complex human motions or physical phenomena, as shown in Fig. 1, 4, 5, and 6. The key reason is that there are no negative training data to provide contrastive signals discouraging physically inconsistent generations.

The emergence of direct preference optimization [58] (DPO) may provide a potential solution to address the limitations of SFT. However, directly using

DPO may encounter three problems: (i) Lack of training data pairs that comprehensively capture physical activities, interactions, and phenomena. (ii) The correct physics preference optimization cannot be guaranteed because vanilla DPO uses generated video as winning case and the physical realism of the generated video is limited and some challenging cases remain difficult to generate with correct physics. The supervision is mainly based on the Bradley–Terry (BT) probabilistic model, which only compares a single pair of generated samples. Such single binary comparison struggles to capture inherently holistic global preference signals of physical plausibility. (iii) Vanilla DPO copies the full model as reference, which consumes substantial GPU memory and decreases efficiency.

To address these issues, we firstly propose a data construction pipeline, PhyAugPipe, that exploits a vision-language model (VLM) to filter text-video data pairs capturing rich physical interactions and phenomena from a large T2V data pool by parsing the entities and reasoning their actions with our designed chain-of-thought (CoT) [73] rule. We use PhyAugPipe to collect a training dataset, PhyVidGen-135K, containing 135K text-video pairs. Secondly, we formulate a novel Physics-aware Groupwise Direct Preference Optimization (PhyGDPO) framework for physically consistent T2V generation. PhyGDPO uses the real-world video that generally follows physical laws as the winning case to guarantee correct physics learning. Different from vanilla pairwise DPO, PhyGDPO is based on the groupwise Plackett–Luce (PL) model, which captures the probability distribution over a group of candidate videos, enabling holistic preference adaptation beyond isolated pairwise comparisons. To further improve physics preference optimization, we propose a Physics-Guided Rewarding (PGR) scheme. PGR leverages a physics-aware VLM to guide data sampling and DPO training to focus on challenging actions and allows physics-violating samples to exert stronger influence. To improve DPO training efficiency and stability, we also propose a LoRA [29]-Switch Reference (LoRA-SR) scheme that does not need to copy the full model as reference like vanilla DPO, which occupies redundant GPU memory. Instead, we freeze the base model as reference and attach LoRA with an environment manager to flexibly switch between reference and action modes. Benefit from our collected dataset and designed techniques, PhyGDPO boosts the physical plausibility of the base T2V model and yields better results than the closed-source models OpenAI Sora2 [7] and Google Veo3.1 [20] on some challenging actions, as shown in Fig. 1, 4, 5, and 6.

In a nutshell, our contributions can be summarized as follows:

- We formulate a principled DPO framework, PhyGDPO, to capture holistic physics advantage signal for physically consistent text-to-video generation.
- We design PGR to guide data sampling and preference optimization to focus on challenging physics. We propose LoRA-SR for efficient DPO training.
- We present PhyAugPipe to construct physics-rich training data pairs. We use PhyAugPipe to collect a large training dataset, PhyVidGen-135K.
- Our PhyGDPO surpasses SOTA methods on PhyGenBench and VideoPhy2 while yielding higher human preference in the user study of physical realism.

## 2 Related Work

### 2.1 Text-to-Video Generation Models

T2V generation [6, 7, 19, 24, 26–28, 32, 60, 61, 70, 84, 89] has witnessed significant progress. Many existing works follow DiT [54] to use a Transformer [63] to predict the noise in diffusion [9, 22, 25, 37, 39, 46–48, 56] or estimate the velocity in flow matching [8, 14, 19, 23, 32, 33, 36, 65–67, 83, 85–87]. Although high visual quality is achieved, accurately modeling the underlying physics-related effects remains challenging. Two recent works, DiffPhy [82] and PhyT2V [77], use an LLM to extend the prompt with explicit physics phenomena, and then iteratively generate the video or finetune a T2V model with extended prompts. Yet, these prompt extension-based methods are easily misled by the mistakes of LLM.

### 2.2 Direct Preference Optimization

Direct Preference Optimization (DPO) [58] is proposed to align LLMs with human preferences and has been adopted in image [18, 35, 51, 64] and video [13, 30, 31, 41–43, 50, 72, 76, 80] generation. For example, DiffusionDPO [64] finetunes a text-to-image diffusion model with human-annotated preference image pairs. VideoDPO [43] adapts DiffusionDPO into T2V generation. However, these DPO frameworks mainly improve the visual quality and aesthetics, while leaving the physics plausibility problem under-explored. We aim to fill this research gap.

## 3 Method

### 3.1 Physics-Augmented Video Data Construction

Due to the scarcity of text-video data pairs with rich physics interactions and phenomena, we design a Physics-Augmented video data construction Pipeline (PhyAugPipe). As shown in Fig. 2, we first use a VLM, Qwen2.5-72B-Instruct [78], equipped with our chain-of-thought (CoT) reasoning in (b) to derive physics-rich videos in (c) by filtering from a large T2V data pool in (a). Then in (d), we cluster the filtered data according to their action categories and sample the data with the rewards of a physics-aware VLM, VideoCon-Physics [4], in (e).

**Data Filtering with CoT.** As shown in Fig. 2 (b), Qwen2.5 with our CoT first parses the elements from the given prompts and video frames, including the physical objects with materials, actions, and forces between them. Based on the parsed elements, Qwen2.5 reasons how the entities interact and what results, and rates the physics richness of the data with a score from 0 to 1. Eventually, Qwen2.5 extends the prompt with explicit physics phenomena. Although we explore the implicit physics reasoning ability of T2V models, our dataset still includes the extended prompts to support other research purposes, such as LLM-guided T2V. Please refer to the supplementary for the details of our CoT. Then in Fig. 2 (c), we threshold the data according to the physics richness.

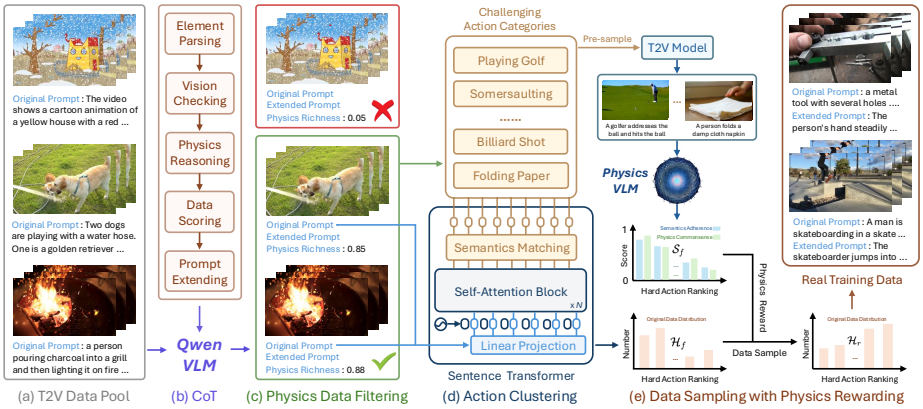


**Fig. 2:** Our Physics-Augmented video data construction Pipeline (PhyAugPipe) uses a VLM, Qwen2.5-72B-Instruct [78], following our designed chain-of-thought (CoT) rule in (b) to select text-video pairs that contain rich physics interactions and phenomena from a large-scale high-quality text-video data pool in (a). Then in (d), we perform action clustering on the filtered data pairs from (c) through the semantics matching via a sentence Transformer [59]. Subsequently, in (e), we use a physics-aware VLM, VideoCon-Physics [4], to evaluate the difficulty of different action categories and then sample the text-video data pairs accordingly as the winning cases of training data.

**Action Clustering via Semantics Matching.** After data filtering, the retained samples may exhibit distributional imbalance on different actions. To check this issue, we categorize the filtered data into semantically coherent action clusters. Specifically, in Fig. 2 (d), we first compile a list containing $K_a$ challenging action categories summarized from the failure cases of frontier video generation models. Then we feed the original text prompt of each filtered sample and the action list into a sentence Transformer [59] to perform fuzzy semantics matching, thereby determining the action category to which each data sample belongs. We count the number of data samples in each action category to obtain the physics action distribution histogram $\mathcal{H}_f$ for the filtered data.

**Data Sampling with Physics Rewarding.** As different action categories vary significantly in physical complexity thus leading to different generation difficulties for the T2V model, we propose to balance the training data accordingly. In particular, we first evaluate how well the T2V model performs on each action category. Based on the performance distribution, we adjust the sampling ratio by allocating more samples to the categories where the model performs poorly. As shown in Fig. 2 (d), we first select the top-$n_c$ samples with the highest semantics matching scores within each category as its representatives. Then we employ a physics-aware VLM, VideoCon-Physics [4], to evaluate each video representative. VideoCon-Physics outputs a semantics adherence score and a physics common-

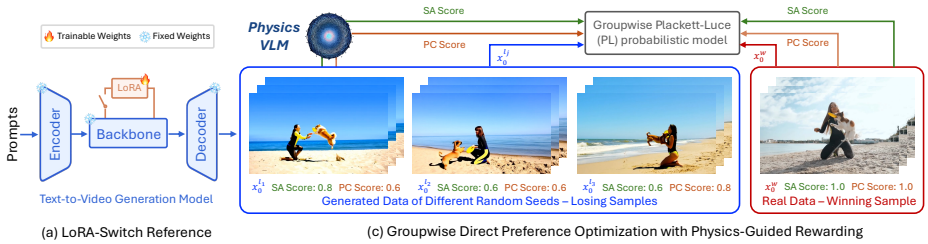


**Fig. 3:** Overview of our PhyGDPO framework. (a) The LoRA-Switch Reference scheme can flexibly toggle between the reference and action modes to efficiently compute DPO advantage and increase the training stability. (b) Our DPO framework is based on the groupwise Plackett-Luce (PL) probabilistic model and uses a physics-aware VLM, VideoCon-Physics [4], to reward the DPO training. The winning case is a real-world video because it captures correct physics as a guarantee for preference learning.

sense score $\in [0, 1]$. We compute the mean score of all representatives of each category to derive the score distribution histogram $\mathcal{S}_f$. Then we define the difficulty of the $k$-th action category as $d_k = 1 - \mathcal{S}_f(k)$ and assign a sampling weight as $r_k = \exp(\tau d_k)$, where $\tau$ controls how the sampling favors challenging actions. Given the total sampling budget $N$, the number of data pairs sampled from the $k$-th action category is $\mathcal{H}_r(k) = \min\left(\mathcal{H}_f(k),\, N \cdot \frac{r_k}{\sum_{j=1}^{K_a} r_j}\right)$, where $\mathcal{H}_r$ denotes the data distribution histogram after rewarding. This sampling allocates more data to challenging actions, encouraging the model to learn more complex physics.

Please note that PhyAugPipe is a filtering-based data pipeline. The VLMs are used to select physics-rich scenes and hard actions from the real-world videos, which always follow the physical laws. Thus, the VLM rewarding only impacts physics richness rather than correctness. We use the original human-written prompts instead of the VLM rewritten ones to avoid the potential bias of VLMs.

### 3.2 Physics-Aware Groupwise Direct Preference Optimization

Existing DPO algorithms compare pairwise data samples, showing limitations in capturing global and holistic preference signals and aligning with human feedback. Besides, vanilla DPO uses generated video as winning case but the physical realism of the generated video is limited and some challenging physics actions are difficult to generate. To address these issues, we formulate a Physics-Aware Groupwise Direct Preference Optimization (PhyGDPO), as illustrated in Fig. 3.

**Groupwise Probability.** We denote the reward as $r(c, x_0)$ with the generation $x_0$ and condition $c$. Different from the normal DPO, we start from the groupwise Plackett-Luce (PL) probabilistic model, as shown in Fig. 3 (b). We use the real video as the winning case $x_0^w$ because it generally follows physical laws, which guarantees correct physics learning, and a set of generated videos as losing cases

$\mathcal{G}^l(c) = \{x_0^{l_1}, \ldots, x_0^{l_m}\}$. The preference probability of the groupwise PL model is

$$p_{\mathrm{PL}}\big(x_0^w \mid \mathcal{G}^l(c), c\big) = \frac{\exp\big(r(c, x_0^w)\big)}{\exp\big(r(c, x_0^w)\big) + \sum_{j=1}^m \exp\big(r(c, x_0^{l_j})\big)}. \tag{1}$$

The reward function $r(c, x_0)$ can be parameterized by a neural network $\phi$ [58] and estimated via minimum negative log-likelihood training as

$$\mathcal{L}_{\mathrm{PL}}(\phi) = -\,\mathbb{E}_{c,\,\mathcal{G}^l(c)}\left[\log \frac{\exp\big(r_\phi(c, x_0^w)\big)}{\exp\big(r_\phi(c, x_0^w)\big) + \sum_{j=1}^m \exp\big(r_\phi(c, x_0^{l_j})\big)}\right]. \tag{2}$$

Vanilla DPO [43, 64] represents $r(c, x_0)$ by the T2V model itself as

$$r(c, x_0) = \beta \log \frac{p_\theta^*(x_0|c)}{p_\psi(x_0|c)} + \beta \log Z(c), \tag{3}$$

where $p_\theta$ and $p_\psi$ denote the conditional probabilities of the trained action model $\theta$ and the fixed reference model $\psi$. $p_\theta^*$ and $Z(c)$ are the unique global optimal solution and the partition function. We plug Eq. (3) into Eq. (2) and cancel the constant term $\beta \log Z(c)$ to derive the groupwise DPO (GDPO) loss as

$$\begin{aligned}\mathcal{L}_{\mathrm{GDPO}}(\theta) &= -\,\mathbb{E}_{c,\,\mathcal{G}^l(c)}\left[\log \frac{\exp\big(\beta f_\theta(x_0^w, c)\big)}{\exp\big(\beta f_\theta(x_0^w, c)\big) + \sum_{j=1}^m \exp\big(\beta f_\theta(x_0^{l_j}, c)\big)}\right] \\ &= \mathbb{E}_{c,\,\mathcal{G}^l(c)}\left[\log\Big(1 + \sum_{j=1}^m \exp\big(\beta(f_\theta(x_0^{l_j}, c) - f_\theta(x_0^w, c))\big)\Big)\right],\end{aligned} \tag{4}$$

where we define the function $f_\theta(x_0, c)$ as the log-likelihood ratio as

$$f_\theta(x_0, c) = \log \frac{p_\theta(x_0|c)}{p_\psi(x_0|c)} = \sum_{k=1}^T \log \frac{p_\theta(x_{t_{k-1}} \mid x_{t_k}, c)}{p_\psi(x_{t_{k-1}} \mid x_{t_k}, c)} \triangleq \sum_{k=1}^T \Delta_k. \tag{5}$$

Here the terminal distribution is inherently represented as the joint transition process over multiple timesteps $t_k$ by the definition of diffusion or flow matching models. Subsequently, we reformulate Eq. (4) and estimate its upper bound using Jensen's inequality into a single timestep for more efficient training as

$$\begin{aligned}\mathcal{L}_{\mathrm{GDPO}}(\theta) &= \mathbb{E}_{c,\,\mathcal{G}^l(c)}\left[\log\Big(1 + \sum_{j=1}^m \exp\big(\beta T \mathbb{E}_k[\Delta_k^{l_j} - \Delta_k^w]\big)\Big)\right] \\ &\le \mathbb{E}_{c,\mathcal{G}^l(c),k}\left[\log\Big(1 + \sum_{j=1}^m \exp\big(\beta T[\Delta_k^{l_j} - \Delta_k^w]\big)\Big)\right].\end{aligned} \tag{6}$$

Although the upper bound in Eq. (6) allows single-timestep training, it still needs to infer the model $2m+2$ times for each group, which requires a long time and large GPU memory. To handle this issue, we exploit an inequality as

$$1 + \sum_{j=1}^m e^{x_j} \le \prod_{j=1}^m \left(1 + e^{\alpha_j x_j}\right)^{\gamma_j}, \quad 0 < \alpha_j \le 1,\ \gamma_j \ge 1/\alpha_j. \tag{7}$$

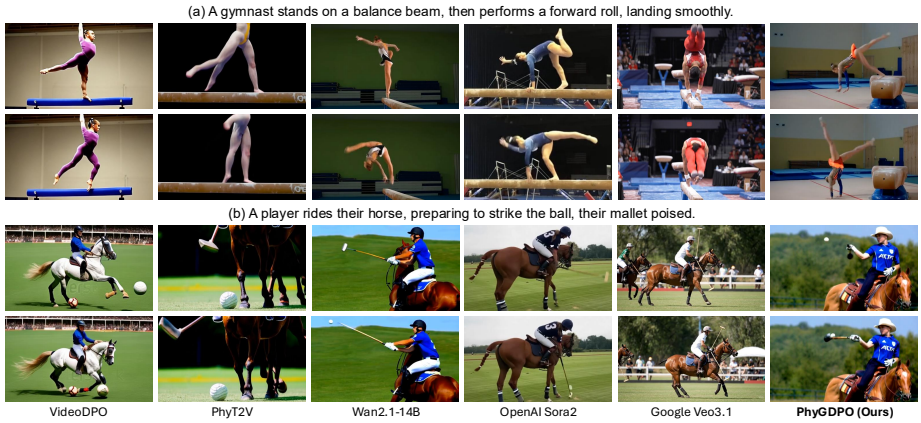


**Fig. 4:** Visual results of two challenging actions (gymnastics and polo) on VideoPhy2. Our method generates more physically consistent videos, showing coherent, deformation-free gymnastic movements and realistic ball–mallet striking interactions.

Subsequently, we can further formulate the upper bound of $\mathcal{L}_{\text{GDPO}}(\theta)$ via Ineq. (7) into a single data sample within each group for more efficient training:

$$\begin{aligned}\mathcal{L}_{\text{GDPO}}(\theta) &\leq \mathbb{E}_{c,\mathcal{G}^l(c),k}\Big[\sum_{j=1}^{m}\gamma_j \log\big(1+\exp(-\alpha_j\beta T[\Delta_k^w-\Delta_k^{l_j}])\big)\Big] \\ &= \mathbb{E}_{c,\mathcal{G}^l(c),k,j}\Big[-m\gamma_j\log\sigma\big(\alpha_j\beta T(\Delta_k^w-\Delta_k^{l_j})\big)\Big].\end{aligned} \tag{8}$$

**Physics-Guided Rewarding.** To further improve physics preference optimization, we design a Physics-Guided Rewarding (PGR) to direct DPO training to focus on challenging physics. Specifically, we parameterize $\gamma_j$ and $\alpha_j$ in Eq. (8) by the normalized semantics adherence score $s_j^{sa} \in [0,1]$ and physics commonsense score $s_j^{pc} \in [0,1]$ predicted by a physics-aware VLM, VideoCon-Physics, as

$$\begin{aligned} v_j &= 1-\frac{s_j^{sa}+s_j^{pc}}{2}, \quad \gamma_j = \frac{1+\lambda\cdot\sigma\big(\kappa_\gamma(v_j-b_\gamma)\big)}{\alpha_{\min}}, \\ \alpha_j &= \alpha_{\min}+(1-\alpha_{\min})\cdot\sigma\big(\kappa_\alpha(v_j-b_\alpha)\big), \end{aligned} \tag{9}$$

where $v_j$ measures the physics difficulty and mainly modulates $\gamma_j$ and $\alpha_j$, $\alpha_{\min}$ sets a lower bound to avoid vanishing gradients, and the sigmoid function $\sigma(\cdot)$ ensures a bounded and smooth adjustment of $\gamma_j$, stabilizing the optimization. $\alpha_j$ adaptively adjusts the sharpness of the preference comparison, while $\gamma_j$ amplifies the physics-guided reward for losing data samples with lower physical plausibility. They are controlled by the hyperparameters $\lambda$, $\kappa_{\alpha,\gamma}$, $b_{\alpha,\gamma}$, and $\alpha_{\min}$.

**Flow Matching Probabilistic Modeling.** Our method is based on the standard rectified flow matching with the sampling formulation as

$$x_t = (1-t)\,x_0 + t\,x_1, \quad x_1 \sim \mathcal{N}(0, I), \tag{10}$$

which induces the oracle velocity $u^*(x_t, t \mid c) = x_1 - x_0$. Discretizing time with step size $h$, the backward update is $x_{t-h} = x_t - h\,u^*(x_t, t \mid c)$. Since the rectified flow one-step reverse transition has no intermediate noise, we follow a vanishing-noise regularization and approximate $p_\theta(x_{t-h} \mid x_t, c)$ and $p_\psi(x_{t-h} \mid x_t, c)$ as

$$\begin{aligned} p_\theta(x_{t-h} \mid x_t, c) &\approx \mathcal{N}\big(x_t - h\,v_\theta(x_t, t, c),\, \varepsilon I\big), \\ p_\psi(x_{t-h} \mid x_t, c) &\approx \mathcal{N}\big(x_t - h\,v_\psi(x_t, t, c),\, \varepsilon I\big), \end{aligned} \tag{11}$$

then let $\varepsilon \to 0$ and use the approximation of the log-ratio between Gaussians with the identity $(x_{t-h} - x_t)/h = -u^*(x_t, t \mid c)$, we can obtain

$$\begin{aligned} \log \frac{p_\theta(x_{t-h} \mid x_t, c)}{p_\psi(x_{t-h} \mid x_t, c)} &\approx -\frac{h^2}{2\varepsilon}\Big(\ell_\theta(x_t, t) - \ell_\psi(x_t, t)\Big), \\ \ell_\theta(x_t, t) &= \|v_\theta(x_t, t, c) - u^*(x_t, t \mid c)\|_2^2, \\ \ell_\psi(x_t, t) &= \|v_\psi(x_t, t, c) - u^*(x_t, t \mid c)\|_2^2, \end{aligned} \tag{12}$$

and reformulate the upper bound of Eq. (8) into the final training objective as

$$\mathcal{L} = \mathbb{E}_{c, \mathcal{G}^l(c), k, j}\Big[ -m\gamma_j \log \sigma\big( -\alpha_j \beta T[(\ell_\theta^w - \ell_\psi^w) - (\ell_\theta^{l_j} - \ell_\psi^{l_j})]\big)\Big], \tag{13}$$

where any global constant scale factor such as $h^2/2\varepsilon$ in Eq. (12) can be absorbed into the hyperparameter $\beta$. For simplicity, we denote $\ell_\theta^w = \ell_\theta(x_{t_k}^w \mid t_k)$, $\ell_\psi^w = \ell_\psi(x_{t_k}^w \mid t_k)$, $\ell_\theta^{l_j} = \ell_\theta(x_{t_k}^{l_j} \mid t_k)$, and $\ell_\psi^{l_j} = \ell_\psi(x_{t_k}^{l_j} \mid t_k)$ in Eq. (13).

**LoRA-Switch Reference.** Previous DPO methods usually copy the full model and fix it as the reference, which occupies redundant GPU memory and leads to unstable and less effective DPO training. To address this issue, we design a LoRA-Switch Reference (LoRA-SR) scheme. As shown in Fig. 3 (b), we freeze the backbone as the reference model $\psi$ and attach trainable LoRA modules to $\psi$ as the action model $\theta$. Specifically, we attach LoRA to the query, key, value, and output linear projection layers of each self-attention in the Transformer backbone and implement an environment manager as the switch to toggle between the reference and action modes. Given the input tokens of the linear layer where the LoRA is attached to as $\mathbf{X} \in \mathbb{R}^{B \times L \times C_{in}}$, our LoRA-SR scheme is formulated as

$$\mathbf{Y} = \mathbf{X}(\mathbf{W} + \mathbf{1}_{action} \cdot \Delta\mathbf{W})^\top = \mathbf{X}(\mathbf{W} + \mathbf{1}_{action} \cdot \frac{\alpha}{r}\mathbf{B}\mathbf{A})^\top, \tag{14}$$

where $\mathbf{W} \in \mathbb{R}^{C_{out} \times C_{in}}$ denotes the frozen weight of the linear layer, $\mathbf{1}_{action} \in \{0, 1\}$ is a binary indicator governed by the environment manager, $\alpha$ is the scaling factor, $r$ is the rank, and $\mathbf{B} \in \mathbb{R}^{C_{out} \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times C_{in}}$ are the learnable LoRA


**Fig. 5:** Comparison on wild user-input actions. Our method generates more physically plausible videos, capturing racket–ball interactions and well-coordinated body motion.

parameters. Our LoRA-SR can flexibly switch the LoRA parameters for the action and reference models to compute the DPO advantage without storing or loading an extra full model. This significantly increases computational efficiency, improves the scalability of model size and stabilizes the DPO training process.

## 4 Experiments

**Dataset.** We use PhyAugPipe to process one million high-quality text-video pairs. After physics data filtering in Fig. 2 (c), we obtain 135K data pairs with sufficient physics interaction and phenomena. Then we perform action clustering and data resampling with physics rewarding to derive 17K data pairs from the 135K filtered data samples. Subsequently, we use the T2V model Wan2.1-14B to generate videos for the 17K text prompts with different random seeds and use VideoCon-Physics [4] to score them. We use the short unextended prompts from the two datasets, VideoPhy2 [5] and PhyGenBench [49], for evaluation.

**Evaluation Metrics.** VideoPhy2 uses another VLM - VideoPhy2-AutoRater [5], which is different from VideoCon-Physics [4] in scoring scheme and training data, to evaluate the semantics adherence and physics commonsense. PhyGenBench evaluation is based on GPT-4o [1], CLIP [57], InternVideo2 [71], and VideoLLaVA [40]. It first detects key physical phenomena and then assesses their order and naturalness, covering 27 physical laws in 4 basic physics domains.

**Implementation Details.** We implement PhyGDPO by pytorch [53] based on a foundation T2V model Wan2.1-14B. Our model is finetuned for 10K steps in

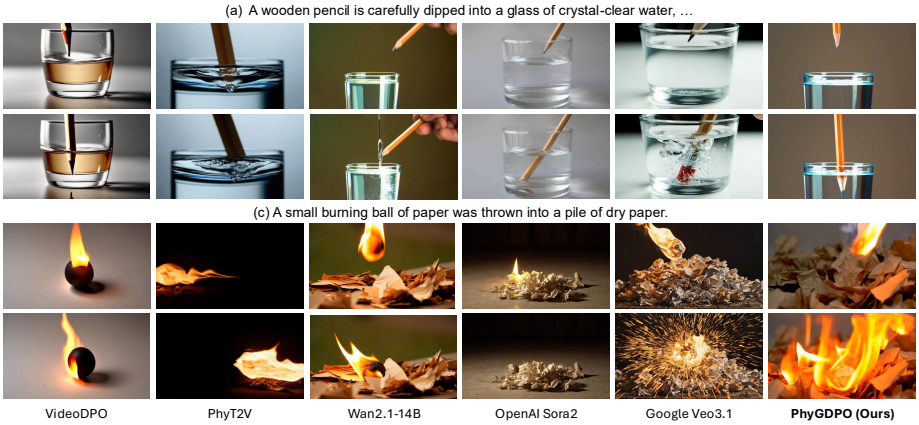


**Fig. 6:** Visual results of two challenging physics phenomena (water refraction and paper combustion) on the PhyGenBench [49] dataset. Our method can model the magnifying convex-lens effect and the realistic light refraction caused by the curved water surface, as well as the physically correct ignition and flame propagation on the paper.

total at a batch size of 8 on 8 H100 GPUs for 6 days. To save GPU memory, we use mixed-precision training [52] with BF16 and sublinear memory training [15]. We use the AdamW optimizer [45] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a weight decay of 0.01. The learning rate is initially set as $1e^{-5}$ and decays to $1e^{-6}$ using cosine annealing [44] algorithm. The training and inference spatial resolution of the video is 480×832. We set $m = 3$, $\tau = 3$, $N = 20000$, $\alpha_{\min} = 0.5$, $k_\gamma = 2.0$, $b_\gamma = 0.4$, $\lambda = 0.6$, $k_\alpha = 5.0$, $b_\alpha = 0.5$. The scale factor and rank of LoRA are 48.

### 4.1 Comparison with State-of-the-Art Methods

**Quantitative Results.** We compare PhyGDPO with state-of-the-art (SOTA) methods in Tab. 1a and 1b. **(i)** Our method surpasses the two recent strongest closed-source commercial models, Sora2 [7] and Veo3.1 [20] across the tracks of hard actions, activities and sports, and the overall score on VideoPhy2. **(ii)** Compared to the SOTA DPO algorithm for video generation (VideoDPO) and SOTA method for physically consistent video generation (PhyT2V), our method surpasses VideoDPO and PhyT2V by 200% and 29% on hard action score of VideoPhy2. Plus, our method is 22%/15% and 35%/23% higher than PhyT2V/VideoDPO on the mechanics and thermal tracks of PhyGenBench.

**User Study.** Besides VLM evaluation, we also conduct a user study with 104 participants with solid knowledge in physics and vision. Each participant is asked to pick up the video that better follows the physical laws. Tab. 1c reports the user preference (%) of our method over competing entries. Each participant

| Methods | Hard | Activity | Interaction | Overall |
|---|---|---|---|---|
| Vcrafter2 [12] | 0.0222 | 0.1071 | 0.0943 | 0.1034 |
| Wan2.1-14B [65] | 0.0111 | 0.1190 | 0.1572 | 0.1288 |
| Hunyuan [38] | 0.0222 | 0.1286 | 0.1572 | 0.1356 |
| VideoDPO [43] | 0.0167 | 0.1310 | 0.1572 | 0.1373 |
| PhyT2V [77] | 0.0389 | 0.1405 | 0.1698 | 0.1492 |
| Sora2 [7] | 0.0389 | 0.1429 | 0.1698 | 0.1508 |
| Veo3.1 [20] | 0.0444 | 0.1405 | 0.1887 | 0.1525 |
| PhyGDPO (Ours) | 0.0500 | 0.1571 | 0.1761 | 0.1627 |

**(a)** Results on VideoPhy2

| Methods | Mechanics | Optics | Thermal | Material | Avg |
|---|---|---|---|---|---|
| Lavie [70] | 0.40 | 0.44 | 0.38 | 0.32 | 0.36 |
| Wan2.1-14B [65] | 0.36 | 0.53 | 0.36 | 0.33 | 0.40 |
| Open-Sora [88] | 0.43 | 0.50 | 0.44 | 0.37 | 0.44 |
| Pika [55] | 0.35 | 0.56 | 0.43 | 0.39 | 0.44 |
| Vchitect-2.0 [21] | 0.41 | 0.56 | 0.44 | 0.37 | 0.45 |
| PhyT2V [77] | 0.45 | 0.55 | 0.43 | 0.53 | 0.50 |
| VideoDPO [43] | 0.48 | 0.60 | 0.47 | 0.58 | 0.54 |
| PhyGDPO (Ours) | 0.55 | 0.60 | 0.58 | 0.47 | 0.55 |

**(b)** Results on PhyGenBench

| Compared Methods | Preference of Ours |
|---|---|
| Vcrafter2 [12] | 94.2 % |
| VideoDPO [43] | 89.4 % |
| PhyT2V [77] | 88.5 % |
| Wan2.1-14B [65] | 86.5 % |
| Hunyuan [38] | 82.7 % |
| Sora2 [7] | 67.3 % |
| Veo3.1 [20] | 64.4 % |
| PhyGDPO (Ours) | – |

**(c)** User Preference

**Table 1:** Quantitative comparison and user preference (%) of our method on the VideoPhy2 [5] and PhyGenBench [49] datasets. Our method achieves better results.

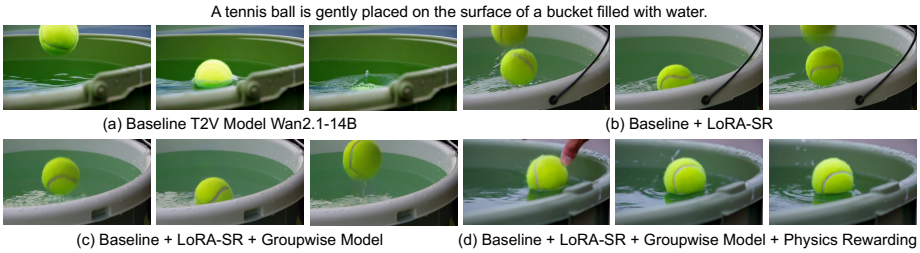


**Fig. 7:** Break-down visual analysis of PhyGDPO. When we progressively apply LoRA-SR, groupwise DPO modeling, and physics rewarding to Wan2.1-14B, the tennis ball in the generated video increasingly conforms to the physical law of buoyancy, floating stably on the water surface instead of sinking or bouncing unrealistically above it.

completes 48 trials, with 24 prompts randomly sampled from VideoPhy2 and 24 from PhyGenBench. Each trial randomly selects one of the 8 baselines in Tab. 1c and compares it to our method given the same prompt. All results in Tab. 1c have 95% confidence intervals no wider than ±2.4%. PhyGDPO is consistently preferred by human evaluators than SOTA methods.

**Qualitative Results.** The visual comparisons on challenging physical actions or phenomena including gymnastics, soccer, basketball, glass smashing, polo, squash, handspring, refraction, and combustion are shown in Fig. 1, 4, 5, and 6. Our method generates videos with more realistic dynamics, undistorted human body, and accurate object interactions, capturing phenomena such as force transfer, material deformation, light refraction, and flame propagation. PhyGDPO shows superior generalization from human activities to complex physical events.

### 4.2 Ablation Study

**Ablation of PhyAugPipe.** We conduct a break-down ablation to study the effect of each component of PhyAugPipe in Tab. 2a. The baseline is directly training PhyGDPO on random text-video data. When we keep the number of

| Method | Hard | Activity | Interaction | Overall |
|---|---|---|---|---|
| Baseline | 0.0333 | 0.1405 | 0.1635 | 0.1475 |
| + Chain-of-Thought | 0.0389 | 0.1452 | 0.1698 | 0.1525 |
| + Action Clustering | 0.0500 | 0.1524 | 0.1698 | 0.1575 |
| + Physics Rewarding | 0.0500 | 0.1571 | 0.1761 | 0.1627 |

**(a)** Break-down ablation of PhyAugPipe

| Method | Hard | Activity | Interaction | Overall |
|---|---|---|---|---|
| Baseline | 0.0111 | 0.1190 | 0.1572 | 0.1288 |
| + LoRA-SR | 0.0278 | 0.1381 | 0.1635 | 0.1458 |
| + Groupwise Model | 0.0389 | 0.1500 | 0.1698 | 0.1559 |
| + Physics Rewarding | 0.0500 | 0.1571 | 0.1761 | 0.1627 |

**(b)** Break-down ablation of PhyGDPO

| Method | Hard | Activity | Interaction | Overall |
|---|---|---|---|---|
| Baseline | 0.0118 | 0.1136 | 0.1447 | 0.1232 |
| Flow-DPO [42] | 0.0296 | 0.1247 | 0.1342 | 0.1283 |
| VideoDPO [43] | 0.0278 | 0.1262 | 0.1509 | 0.1305 |
| PhyGDPO (Ours) | 0.0444 | 0.1357 | 0.1635 | 0.1407 |

**(c)** Comparison with SOTA DPO methods

| Method | GPU Memory | Storage Space | Hard Score | Overall Score |
|---|---|---|---|---|
| Baseline | - | - | 0.0118 | 0.1232 |
| LoRA-SFT | 24.7GB | 84MB | 0.0167 | 0.1283 |
| w/o LoRA-SR | 48.7GB | 5.3GB | 0.0389 | 0.1373 |
| with LoRA-SR | 25.3GB | 84MB | 0.0444 | 0.1407 |

**(d)** Ablation of LoRA-SR scheme

| Method | PhyT2V | VideoDPO | Wan2.1-14B | Veo3.1 | Sora2 | PhyGDPO |
|---|---|---|---|---|---|---|
| Overall | 0.3186 | 0.3288 | 0.4119 | 0.5034 | 0.5373 | 0.5525 |

**(e)** Cross-VLM evaluation with Gemini-2.5-pro

| Method | Vcrafter2 | + Flow-DPO | + VideoDPO | + PhyGDPO |
|---|---|---|---|---|
| Overall | 0.1034 | 0.1128 | 0.1373 | 0.1426 |

**(f)** Cross-T2V-model evaluation

**Table 2:** Ablation on VideoPhy2 [5]. (a) and (b) study the components of PhyAugPipe and PhyGDPO towards better performance on Wan2.1-14B. (c) compares VideoDPO with PhyGDPO under the same settings and (d) studies LoRA-SR on Wan2.1-1.3B. (e) and (f) are cross evaluation by changing the VLM scorer and T2V base model.

training data the same and apply the data filtering with CoT, action clustering with semantics matching, and data sampling with physics rewarding, all track scores are improved. Especially the score on the hard actions gains by over 50%. These results suggest the effectiveness of our data construction techniques.

**Ablation of PhyGDPO.** Tab. 2b shows a break-down ablation for PhyGDPO. When progressively using LoRA-SR, PL groupwise probabilistic model, and physics-guided rewarding, the performance gains by large margins. The score on hard actions yields 4.5× improvement. Besides, we conduct a visual analysis in Fig. 7 for PhyGDPO. When gradually using our techniques, the tennis in the generated video no longer presents ghosting artifacts and its floating motion on the water surface better conforms to the physical laws of fluid buoyancy.

**PhyGDPO *vs.* SOTA DPO.** For fair comparison with two SOTA DPO methods: Flow-DPO [42] and VideoDPO, we use them to finetune Wan2.1-1.3B with the same data and settings. As reported in Tab. 2c, PhyGDPO surpasses Flow-DPO and VideoDPO across all tracks by large margins, especially on the hard action track, where it yields 50% improvement. We conduct a visual comparison in Fig. 8 on the challenging weightlifting action. Flow-DPO and VideoDPO generate distorted or unstable human body poses. In contrast, PhyGDPO generates coherent human motion with stable shapes and balanced force dynamics. Besides, we also compare with DiffusionNFT [87], which does not model the preference probability but adopts online GRPO reward to reweight the flow-matching loss. It achieves 0.1291/0.0262 in overall/hard scores, lower than PhyGDPO.

| $\alpha_{\min}$ | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
|---|---|---|---|---|---|
| Overall | 0.1582 | 0.1591 | 0.1627 | 0.1615 | 0.1608 |

**(a)** Analysis of $\alpha_{\min}$

| $k_\alpha$ | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 |
|---|---|---|---|---|---|
| Overall | 0.1611 | 0.1579 | 0.1627 | 0.1593 | 0.1618 |

**(b)** Analysis of $k_\alpha$

| $b_\alpha$ | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
|---|---|---|---|---|---|
| Overall | 0.1582 | 0.1579 | 0.1627 | 0.1598 | 0.1602 |

**(c)** Analysis of $b_\alpha$

| $\lambda$ | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
|---|---|---|---|---|---|
| Overall | 0.1592 | 0.1586 | 0.1608 | 0.1627 | 0.1615 |

**(d)** Analysis of $\lambda$

| $k_\gamma$ | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 |
|---|---|---|---|---|---|
| Overall | 0.1588 | 0.1609 | 0.1627 | 0.1592 | 0.1598 |

**(e)** Analysis of $k_\gamma$

| $b_\gamma$ | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
|---|---|---|---|---|---|
| Overall | 0.1586 | 0.1627 | 0.1582 | 0.1607 | 0.1595 |

**(f)** Analysis of $b_\gamma$

**Table 3:** Analysis of the hyperparameters in the physics rewarding of Eq.(9) on VideoPhy2. When tweaking a parameter, other parameters are fixed at the optimal values.

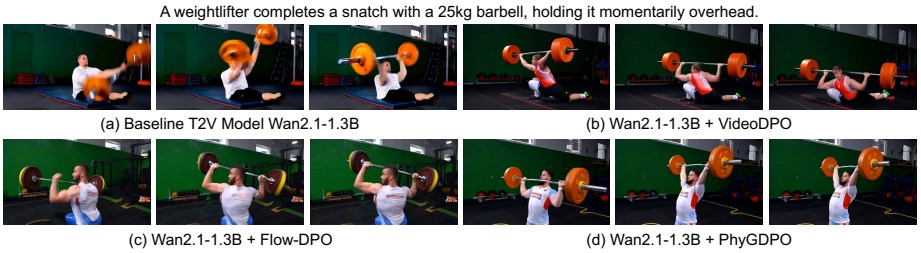


**Fig. 8:** Comparison with state-of-the-art DPO algorithms on Wan2.1-1.3B. Using our PhyGDPO can generate more physically plausible video with undistorted human body.

**Analysis of LoRA-SR.** We conduct an ablation study of LoRA-SR in Tab. 2d. In our PhyGDPO training, using LoRA-SR reduces GPU memory consumption by 44% and achieves over 60× compression in storage size, while still improving the score on hard actions by 14% compared to the version without using LoRA-SR. We also conduct a visual analysis in Fig. 9 (c), using our LoRA-SR in DPO training can generate more plausible physical interactions between human hands and the snow shovel. Besides, we also compare with SFT using the same LoRA in Tab. 2d. PhyGDPO with LoRA-SR uses almost the same GPU memory but surpasses LoRA-SFT by large margins, especially on hard actions.

**Cross Evaluation.** **(i)** To provide stronger evidence that our method improves the physical plausibility, we conduct a cross-VLM evaluation. We replace the VLM scorer of VideoPhy2 by a stronger physics evaluator, Gemini-2.5-pro [17], to score the generated videos with the same protocol. The results are shown in Tab. 2e. Our method still outperforms SOTA T2V models. **(ii)** We change the T2V model to Vcrafter2 [12] and apply different DPO methods on it with the same setting for fair comparison in Tab. 2f. PhyGDPO consistently surpasses SOTA DPO methods, suggesting the generalization ability of our method.

**Parameter Analysis.** **(i)** We analyze the effects of the hyperparameters in Eq. (9). When tweaking a hyperparameter, other hyperparameters are fixed at


**Fig. 9:** Visual analysis of our LoRA-SR scheme. Applying our LoRA-SR can generate more accurate and plausible hand–shovel interaction in the snow shoveling scenario.

their optimal values. The results are shown in Tab. 3. The best performance is achieved when $\alpha_{\min} = 0.5$, $k_\alpha = 5.0$, $b_\alpha = 0.5$, $\lambda = 0.6$, $k_\gamma = 2.0$, and $b_\gamma = 0.4$. As shown in Tab. 2a, when we do not use the physics-guided rewarding, the model achieves 0.1575 on the overall score. And in all hyperparameter configurations of Tab. 3, our method consistently achieves performance gains. These results clearly demonstrate the effectiveness and robustness of our physics reward design. **(ii)** We study the effect of group size $m$ in Eq. (13) by gradually increasing it. The performance saturates when group size $m > 3$, while the data construction time and training time increase. Thus, we set $m = 3$ for a trade-off. **(iii)** We compare Jenson approximation with multi-step loss in Ineq. (6) by increasing $T$. The multi-step loss easily suffers from overfitting and unstable training, which dramatically degrades the performance by up to 15%.

**Limitation.** Our method still faces challenges in complex momentum modeling, *e.g.*, having subtle drifts in trajectory or velocity in multi-ball collisions.

## 5 Conclusion

In this paper, we focus on studying a challenging problem, physically consistent T2V generation, without using LLM for prompt extension in inference. Our goal is to explore the implicit physics reasoning ability of the T2V model itself. To this end, we first propose a data construction method, PhyAugPipe, to collect a training dataset containing 135K text-video pairs with rich physics interaction and phenomena. Then we formulate a principled groupwise DPO framework, PhyGDPO, with two technical designs PGR and LoRA-SR. PhyGDPO efficiently post-trains Wan2.1-T2V-14B on our data to boost the physics plausibility in video generation. Comprehensive experiments demonstrate that our method quantitatively and qualitatively outperforms SOTA algorithms and even achieves more physically consistent results than closed-source T2V models. The user study also shows that our method is preferred by human evaluators.

**Acknowledgements:** This work was supported by the office of Naval Research with award N000142412696.

## References


1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
3. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025)
4. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. In: ICLR (2025)
5. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In: ICLR (2026)
6. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
7. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog (2024)
8. Cai, Y., Zhang, H., Chen, X., Xing, J., Hu, Y., Zhou, Y., Zhang, K., Zhang, Z., Kim, S.Y., Wang, T., et al.: Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. In: NeurIPS (2025)
9. Cai, Y., Zhang, H., Zhang, K., Liang, Y., Ren, M., Luan, F., Liu, Q., Kim, S.Y., Zhang, J., Zhang, Z., Zhou, Y., Zhang, Y., Yang, X., Lin, Z., Yuille, A.: Baking gaussian splatting into diffusion denoiser for fast and scalable single-stage image-to-3d generation and reconstruction. In: ICCV (2025)
10. Che, H., He, X., Liu, Q., Jin, C., Chen, H.: Gamegen-x: Interactive open-world game video generation. In: ICLR (2025)
11. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
12. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR (2024)
13. Chen, H.H., Huang, H., Chen, Q., Yang, H., Lim, S.N.: Hierarchical fine-grained preference optimization for physically plausible video generation. In: NeurIPS (2025)
14. Chen, S., Ge, C., Zhang, Y., Zhang, Y., Zhu, F., Yang, H., Hao, H., Wu, H., Lai, Z., Hu, Y., et al.: Goku: Flow based video generative foundation models. In: CVPR (2025)
15. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174 (2016)
16. Chen, X., Guo, J., He, T., Zhang, C., Zhang, P., Yang, D.C., Zhao, L., Bian, J.: Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. arXiv preprint arXiv:2411.00785 (2024)

17. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
18. Croitoru, F.A., Hondru, V., Ionescu, R.T., Sebe, N., Shah, M.: Curriculum direct preference optimization for diffusion and consistency models. In: CVPR (2025)
19. Davtyan, A., Sameni, S., Favaro, P.: Efficient video prediction via sparsely conditioned flow matching. In: ICCV (2023)
20. DeepMind, G.: Veo-3: A scalable and controllable text-to-video model. https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf (2025), technical Report, Google DeepMind
21. Fan, W., Si, C., Song, J., Yang, Z., He, Y., Zhuo, L., Huang, Z., Dong, Z., He, J., Pan, D., et al.: Vchitect-2.0: Parallel transformer for scaling up video diffusion models. arXiv preprint arXiv:2501.08453 (2025)
22. Gao, P., Zhuo, L., Liu, D., Du, R., Luo, X., Qiu, L., Zhang, Y., Lin, C., Huang, R., Geng, S., et al.: Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945 (2024)
23. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
24. Guo, H., Jia, Z., Li, J., Li, B., Cai, Y., Wang, J., Li, Y., Lu, Y.: Efficient autoregressive video diffusion with dummy head. arXiv preprint arXiv:2601.20499 (2026)
25. Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Li, F.F., Essa, I., Jiang, L., Lezama, J.: Photorealistic video generation with diffusion models. In: ECCV (2024)
26. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
27. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
28. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. NeurIPS (2022)
29. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
30. Huang, Z., Zhang, K., Ding, Y., Gao, C., Ding, R., Chen, Y., Zuo, W.: Mind the generative details: Direct localized detail preference optimization for video diffusion models. arXiv preprint arXiv:2601.04068 (2026)
31. Ji, S., Chen, X., Tao, X., Wan, P., Zhao, H.: Physmaster: Mastering physical representation for video generation via reinforcement learning. arXiv preprint arXiv:2510.13809 (2025)
32. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: ICLR (2025)
33. Ju, X., Wang, T., Zhou, Y., Zhang, H., Liu, Q., Zhao, N., Zhang, Z., Li, Y., Cai, Y., Liu, S., et al.: Editverse: Unifying image and video editing and generation with in-context learning. arXiv preprint arXiv:2509.20360 (2025)
34. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. In: ICML (2025)

35. Karthik, S., Coskun, H., Akata, Z., Tulyakov, S., Ren, J., Kag, A.: Scalable ranked preference optimization for text-to-image generation. In: ICCV (2025)
36. Ke, L., Xu, H., Ning, X., Li, Y., Li, J., Li, H., Lin, Y., Jiang, D., Yang, Y., Zhang, L.: Proreflow: Progressive reflow with decomposed velocity. In: CVPR (2025)
37. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. In: ICML (2024)
38. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
39. Li, X., Chu, W., Wu, Y., Yuan, W., Liu, F., Zhang, Q., Li, F., Feng, H., Ding, E., Wang, J.: Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. arXiv preprint arXiv:2309.00398 (2023)
40. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: EMNLP (2024)
41. Liu, J., Li, J., Deng, J., Li, G., Zhou, S., Fang, Z., Lao, S., Deng, Z., Zhu, J., Ma, T., et al.: Dreamontage: Arbitrary frame-guided one-shot video generation. arXiv preprint arXiv:2512.21252 (2025)
42. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Qin, W., Xia, M., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
43. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: CVPR (2025)
44. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2017)
45. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
46. Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P., Ding, M.: Vdt: General-purpose video diffusion transformers via mask modeling. In: ICLR (2024)
47. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. TMLR (2024)
48. Menapace, W., Siarohin, A., Skorokhodov, I., Deyneka, E., Chen, T.S., Kag, A., Fang, Y., Stoliar, A., Ricci, E., Ren, J., et al.: Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In: CVPR (2024)
49. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In: ICML (2025)
50. Mu, L., Wang, Q., Jiang, F., Wang, M., Fan, Y., Xu, M., Zhang, K.: Fantasyhsi: Video-generation-centric 4d human synthesis in any scene through a graph-based multi-agent framework. arXiv preprint arXiv:2509.01232 (2025)
51. Na, S., Kim, Y., Lee, H.: Boost your human image generation model via direct preference optimization. In: CVPR (2025)
52. Narang, S., Diamos, G., Elsen, E., Micikevicius, P., Alben, J., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al.: Mixed precision training. In: ICLR (2018)
53. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
54. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
55. Pika: Pika art. https://pika.art/ (2025), accessed: 2025-05-04

56. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2022)
58. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: NeurIPS (2023)
59. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP (2019)
60. Seedance, T., Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., et al.: Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507 (2025)
61. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. arXiv preprint arXiv:2512.16776 (2025)
62. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. In: ICLR (2025)
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems (2017)
64. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: CVPR (2024)
65. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
66. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025)
67. Wang, J., Ma, A., Cao, K., Zheng, J., Feng, J., Zhang, Z., Pang, W., Liang, X.: Wisa: World simulator assistant for physics-aware text-to-video generation. In: NeurIPS (2026)
68. Wang, Q., Luo, Y., Shi, X., Jia, X., Lu, H., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In: SIGGRAPH (2025)
69. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV (2024)
70. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. IJCV (2025)
71. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: ECCV (2024)
72. Wang, Z., Wang, K., Zhang, L.: Phydetex: Detecting and explaining the physical plausibility of t2v models. arXiv preprint arXiv:2512.01843 (2025)
73. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022)

74. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: CVPR (2024)
75. Wu, W., Liu, M., Zhu, Z., Xia, X., Feng, H., Wang, W., Lin, K.Q., Shen, C., Shou, M.Z.: Moviebench: A hierarchical movie level dataset for long video generation. In: CVPR (2025)
76. Wu, Z., Kag, A., Skorokhodov, I., Menapace, W., Mirzaei, A., Gilitschenski, I., Tulyakov, S., Siarohin, A.: Densedpo: Fine-grained temporal preference optimization for video diffusion models. arXiv preprint arXiv:2506.03517 (2025)
77. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In: CVPR (2025)
78. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Fan, Z.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)
79. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: CVPR (2024)
80. Yang, X., Tan, Z., Li, H.: Ipo: Iterative preference optimization for text-to-video generation. arXiv preprint arXiv:2502.02088 (2025)
81. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
82. Zhang, K., Xiao, C., Mei, Y., Xu, J., Patel, V.M.: Think before you diffuse: Llms-guided physics-aware video generation. arXiv preprint arXiv:2505.21653 (2025)
83. Zhang, S., Li, W., Chen, S., Ge, C., Sun, P., Zhang, Y., Jiang, Y., Yuan, Z., Peng, B., Luo, P.: Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. arXiv preprint arXiv:2502.05179 (2025)
84. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
85. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: Videorepa: Learning physics for video generation through relational alignment with foundation models. In: NeurIPS (2026)
86. Zhang, Y., Yang, H., Zhang, Y., Hu, Y., Zhu, F., Lin, C., Mei, X., Jiang, Y., Peng, B., Yuan, Z.: Waver: Wave your way to lifelike video generation. arXiv preprint arXiv:2508.15761 (2025)
87. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. In: ICLR (2025)
88. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
89. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)