LEARNING PHYSICAL REWARDS FROM VIDEO PREFERENCES FOR REINFORCEMENT LEARNING OF VIDEO GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generators produce visually convincing clips that still violate basic physics, such as a hand passing through a tabletop. Prior reinforcement learning (RL) methods for physical video generation reward the generator with hand-designed verifiers, each checking one property from object tracks, 3D estimates, or simulation states. These measurements are unreliable in open-world videos, and violations of other properties go unpenalized. We instead learn the reward from a physics preference dataset of 10k prompt–video pairs. The prompts cover 25 sub-scenes in five categories: human motion, indoor and everyday objects, industry and robotics, traffic and vehicles, and nature and materials. The label of each pair is known from how the pair is built, e.g., a real clip over a generated one, so annotators only write rationales. A rationale scores each video on six physical dimensions, such as mechanics and geometry, and lists each violation with its time interval and bounding box. On these data we train Phy-Judger, a VLM reward model that writes a rationale before scoring a video, so each score is tied to specific violations. We also train a prompt policy jointly with the video policy; it rewrites each prompt to add missing physical conditions, such as material, contact, and support. Joint RL with Phy-Judger raises the VBench-2.0 Physics score of MiniMax-H3 from 75.80 to 78.83 and its PhyWorldBench physical commonsense score from 0.155 to 0.203. We will release the code, checkpoints and datasets.

1 INTRODUCTION

Modern video generators (Bar-Tal et al., 2024; Yang et al., 2025; Wan Team, 2025) produce visually convincing clips, yet a beautiful appearance does not guarantee a plausible physical event. A hand may pass through a tabletop on contact, oil and water may become a single, stable liquid when poured together, or two people may merge into one as they move close together. Standard denoising and flow-matching objectives do not directly punish these failures, motivating post-training with a physical reward (Figure 1).

Prior methods guide generation with representations of physical dynamics, including physical embeddings (Ji et al., 2025), state-transition tokens (Shang et al., 2026), motion features (Bhowmik et al., 2026), or executable simulation programs (Li et al., 2026a). Others post-train generators on preferences from real videos and automated physical assessments (Cai et al., 2026; Qian et al., 2025; Wang et al., 2025b; Du et al., 2026); direct preference optimization (DPO) may inherit the appearance of preferred videos. Closest to our setting, RL methods use verifiers that each check one property: rigid-body motion (Zhang et al., 2026), simulation-grounded 4D dynamics (Lu et al., 2026), or 3D structure and camera motion (Wang et al., 2026c; Li et al., 2026b; Wang et al., 2026e). These verifiers depend on object tracks, 3D estimates, or simulation states that are often unreliable in open-world videos, and leave violations outside their scope unpenalized.

We learn a physical reward from 10k preference pairs spanning 25 sub-scenes in five categories. These preferences train only the reward; the generator is optimized with online RL. Half of the pairs compare a real clip with a generated one; the other half compare two generated videos of the same prompt, matching the comparisons made during RL and reducing reliance on real video appearance. They pair stronger and weaker generators, or original and physically perturbed prompts

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

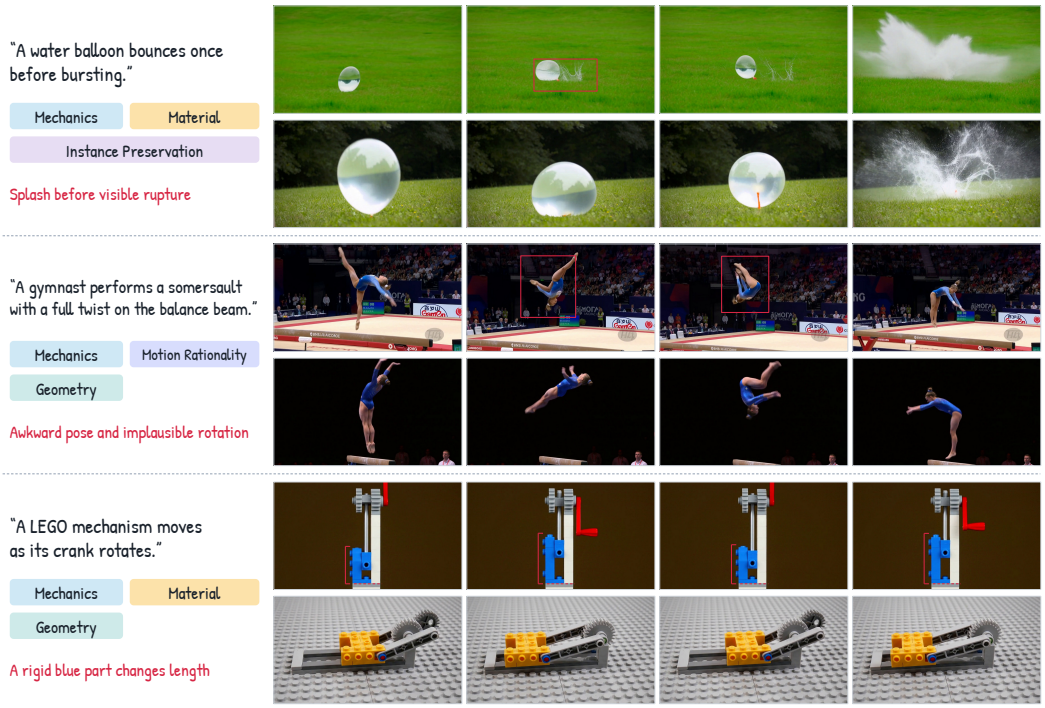


Figure 1: Joint RL with the Phy-Judger reward corrects physical violations across multiple dimensions. For each prompt, the top row is generated before RL, with the violation written and marked in red; the bottom row is generated after joint RL, where the violation is corrected. Tags list the physical dimensions involved. More examples are shown in Figure 6.

rendered with the same generator. Our reward model, *Phy-Judger*, is trained on rationales that localize physical violations and writes this evidence before its preference label, making each score grounded in visible violations.

Physical plausibility is subjective to judge, and annotation criteria vary across scenes. We fix each preference label by pair type, so annotators do not judge which video is better; they only identify the visible violations supporting the label. Human gold rationales for each sub-scene guide VLM annotators through few-shot exemplars. We discard pairs whose label lacks visible support and retain a rationale only when a separate checker can recover the label from it alone.

Underspecified prompts leave the generator to infer material, support, or contact conditions, which can cause physical failures. We therefore jointly train a prompt policy to make these conditions explicit and a video policy to render the requested event.

Our contributions are: (i) a 10k-pair physics preference dataset across 25 sub-scenes, with labels determined by pair type (real vs. generated, stronger vs. weaker generator, or original vs. perturbed prompt) and rationales that localize violations; (ii) *Phy-Judger*, a VLM reward model that generates evidence before its preference label; and (iii) joint RL of prompt and video policies using this reward, in which the prompt policy makes implicit physical conditions explicit.

2 Related Work

2.1 Physical video generation

Video generators are commonly trained with denoising or flow-matching objectives (Ho et al., 2020; Lipman et al., 2023; Liu et al., 2023), and physical benchmarks show that they still violate physical commonsense (Zheng et al., 2025; Zhou et al., 2026; Bansal et al., 2026; Gu et al., 2026; Guo et al., 2025). Prior methods follow three lines of work. The first injects representations of physical

dynamics into generation (Ji et al., 2025; Shang et al., 2026; Lin et al., 2025; Bhowmik et al., 2026; Li et al., 2026a), but does not check the generated video for physical failures. The second post-trains generators with DPO on preferences built from real videos or automated physical assessments (Cai et al., 2026; Qian et al., 2025; Wang et al., 2025b; Du et al., 2026; Kupyn et al., 2025), and may inherit the appearance of the preferred videos. The third, closest to ours, uses RL rewards from verifiers that each check one property, such as rigid-body motion (Zhang et al., 2026), Newtonian dynamics (Le et al., 2025), 4D dynamics (Lu et al., 2026), or 3D structure and camera motion (Wang et al., 2026c; Li et al., 2026b; Wang et al., 2026e). These verifiers rely on measurements that are often unreliable in open-world videos, and leave other violations unpenalized.

Among learned judges, PhyDetEx detects and explains physical violations from real–generated pairs (Wang et al., 2026d), while general video rewards (Liu et al., 2025b; Wang et al., 2026b) are not tailored to physics. We instead learn a single physical reward from preference pairs that include generated–generated comparisons and rationales localizing each violation, and use it for online RL.

2.2 Online reinforcement learning and prompt refinement

RL has become the leading approach for post-training visual generators, usually with general or task-specific rewards (Liu et al., 2025b; Wang et al., 2026c). Online RL methods include DDPO (Black et al., 2024), DPOK (Fan et al., 2023), Flow-GRPO (Liu et al., 2025a), DanceGRPO (Xue et al., 2025), and DiffusionNFT (Zheng et al., 2026). PromptRL optimizes prompt and generator policies together (Wang et al., 2026a), while PhyPrompt trains a physical prompt refiner for a fixed generator (Wu et al., 2026). We use our learned physical reward to jointly optimize the video generator and prompt refinements that specify missing physical conditions while preserving the original entities, actions, and outcome.

3 Preliminaries

Following DDPO (Black et al., 2024) and Flow-GRPO (Liu et al., 2025a), the iterative denoising process is formulated as a Markov decision process with state $s_t \triangleq (c, t, x_t)$, action $a_t \triangleq x_{t-1}$, policy $\pi_\theta(a_t \mid s_t) \triangleq p_\theta(x_{t-1} \mid x_t, c)$, deterministic transitions, and a reward $r(x_0, c)$ given only at the final step. Since the deterministic ODE sampler of flow matching models provides neither exploration nor a tractable $p_\theta(x_{t-1} \mid x_t, c)$, Flow-GRPO converts it into an equivalent SDE that matches the marginal distribution at all timesteps. Its Euler–Maruyama discretization defines an isotropic Gaussian policy $\pi_\theta(x_{t-1} \mid x_t, c)$, where t indexes denoising steps. Given a prompt c, $\pi_{\theta_{\text{old}}}$ samples a group of G videos $\{x_0^i\}_{i=1}^G$, and the advantage of the i-th video is calculated by normalizing the group-level rewards,

$$\hat{A}^{\,i} = \frac{r(x_0^i, c) - \text{mean}\big(\{r(x_0^j, c)\}_{j=1}^G\big)}{\text{std}\big(\{r(x_0^j, c)\}_{j=1}^G\big)}. \tag{1}$$

Flow-GRPO optimizes the policy by maximizing

$$\mathcal{J}_{\text{Flow-GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\min\Big(\rho_t^i(\theta)\,\hat{A}^{\,i},\ \text{clip}\big(\rho_t^i(\theta), 1-\varepsilon, 1+\varepsilon\big)\hat{A}^{\,i}\Big) - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)\Big)\Bigg], \tag{2}$$

where $\rho_t^i(\theta) = p_\theta(x_{t-1}^i \mid x_t^i, c)/p_{\theta_{\text{old}}}(x_{t-1}^i \mid x_t^i, c)$ is the per-step importance ratio.

4 Method

Our framework improves the physical plausibility of a video generator π_θ in three stages that follow the flow of supervision (Figure 2): we first construct a physics preference dataset (Section 4.1), then train a VLM-based physical reward model on it (Section 4.2), and finally use this reward to optimize

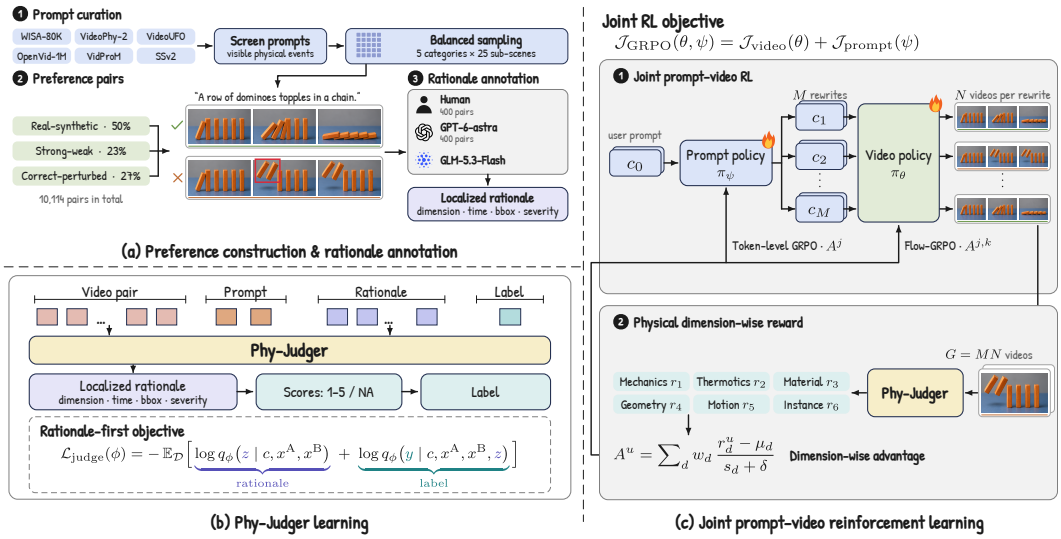


Figure 2: Method overview. **(a)** We build physics preference pairs whose labels are fixed by construction and annotate them with localized rationales. **(b)** Phy-Judger, trained on these data, writes a rationale before predicting the label and scores each applicable physical dimension. **(c)** The prompt and video policies are jointly optimized on shared rollouts scored by Phy-Judger, using dimension-wise advantages.

the generator (Section 4.3). In the last stage, a prompt policy π_ψ is trained jointly with π_θ on the same generated videos, rewriting user prompts to make physical conditions explicit while preserving the requested content.

4.1 Physics Preference Dataset

Data sources and pair construction. We screen six public datasets, WISA-80K (Wang et al., 2025a), VideoPhy-2 (Bansal et al., 2026), VideoUFO (Wang & Yang, 2025), OpenVid-1M (Nan et al., 2025), VidProM (Wang & Yang, 2024), and Something-Something V2 (Goyal et al., 2017), for prompts that describe a visually testable physical event, and group them into 25 sub-scenes under five categories (human motion, indoor and everyday objects, industry and robotics, traffic and vehicles, and nature and materials), which serve as the units for balanced sampling (Appendix A.1). These corpora combine physics-curated prompts (WISA-80K, VideoPhy-2), prompts written by real users (VidProM, VideoUFO), open-domain and everyday object scenes (OpenVid-1M, Something-Something V2). From these prompts, we construct preference pairs $(c, x^{\text{A}}, x^{\text{B}}, y)$, in which both videos share the prompt c and the label $y \in \{\text{A}, \text{B}\}$ marks the more physically plausible one.

We build three types of video pairs, in which the preferred label is known by construction: a real clip is preferred over a video generated from its caption (*real–synthetic*), a stronger generator over a weaker one (*strong–weak*, e.g., Wan3.0 or MiniMax-H3 over Wan2.1-1.3B), and a video generated from c over one generated with the same seed from a perturbed prompt $\tilde{c}$ that adds a single violation (*correct–perturbed*). For example, $\tilde{c}$ adds "the chain suddenly stops halfway" to c = "a row of dominoes topples in a chain". Real clips provide examples of physically correct motion, but a judge could separate them from generated videos by appearance alone. The other two types avoid this problem because both videos are generated, just like the videos the judge scores during RL. The detailed sources, filtering, and composition are summarized in Appendix A.2.

We annotate each pair with a rationale z which includes, for each video, per-dimension scores and the visible violations, each with its physical dimension, time interval, key-frame bounding box, and severity. Specific annotation is tiered by cost: humans write 16 rationales per sub-scene (400 in total), GPT-6-astra annotates another 400 pairs with few-shot human exemplars, and GLM-5.3-Flash annotates the rest with exemplars from these 800; each annotator sees c, both videos, and y. A construction label can be wrong, e.g., when both generated videos are equally plausible or equally implausible; the annotator then marks the pair UNSUPPORTED, and it is discarded. Because

an annotator may misstate the evidence, a separate model checker infers a label $\hat{y}$ from z alone, and we retain $\mathcal{D} = \{(c, x^{\mathrm{A}}, x^{\mathrm{B}}, z, y) : \hat{y} = y\}$ among the rationales that pass the schema, interval, and box checks. Rejected pairs are replaced from the same sub-scene and pair type to keep the target composition. In total, we retain 10,114 pairs: 50% real–synthetic, 23% strong–weak, and 27% correct–perturbed.

4.2 VLM PHYSICAL REWARD

The 10k-scale preference dataset provides supervision for ***Phy-Judger***, the VLM-based judge that scores generated videos during RL. We train Phy-Judger q_ϕ as a student model to generate the retained rationale z and the preference label y for each pair video. Following the rationale-first supervision in RationalRewards (Wang et al., 2026b), we minimize

$$\mathcal{L}_{\mathrm{judge}}(\phi) = -\mathbb{E}_{\mathcal{D}}\Big[\log q_\phi\big(z \mid c, x^{\mathrm{A}}, x^{\mathrm{B}}\big) + \log q_\phi\big(y \mid c, x^{\mathrm{A}}, x^{\mathrm{B}}, z\big)\Big]. \tag{3}$$

Training and testing share the same six-dimensional rubric and violation schema; the pointwise and pairwise judge prompts, as well as the score scale, are provided in Appendix A.5. At test time, given a single generated video x, Phy-Judger first generates a rationale z_x and then predicts an ordinal score for each applicable dimension $d \in \mathcal{A}(x, c)$. The pointwise output serves as the physical reward for generator RL.

4.3 JOINT PHYSICAL REINFORCEMENT LEARNING

A user prompt c_0 often leaves physical premises such as material, contact, or support unspecified. We therefore jointly train a prompt policy π_ψ that rewrites c_0 into prompts c_j stating these premises, and the video policy π_θ that renders them. For each c_0, π_ψ samples M rewrites and π_θ generates N videos $\{x_0^{j,k}\}_{k=1}^{N}$ per rewrite c_j ($j = 1, \ldots, M$), where k indexes the videos sharing c_j, giving $G = MN$ videos. Instead of averaging dimension scores into a scalar reward, we keep the Phy-Judger score $r_d^{j,k}$ on each applicable physical dimension d.

Both policies use a dimension-wise advantage for a candidate u with scores r_d^u:

$$A^u = \sum_d w_d \, \frac{r_d^u - \mu_d}{s_d + \delta}, \tag{4}$$

where μ_d is the mean of r_d over candidates with the same prompt as u, s_d is the standard deviation over the whole batch as in the Flow-GRPO implementation, both over candidates to which d applies, the term is zero if d does not apply to u, and $\sum_d w_d = 1$. Normalizing before weighting keeps the most variable dimension from dominating. For π_θ, μ_d averages over the N videos generated from the same rewrite c_j. For π_ψ, μ_d averages over the M rewrites of the same prompt c_0.

The video policy maximizes the Flow-GRPO objective in equation 2 with advantages $A^{j,k}$, denoted $\mathcal{J}_{\mathrm{video}}(\theta)$, and the prompt policy maximizes its token-level GRPO counterpart with advantages A^j, denoted $\mathcal{J}_{\mathrm{prompt}}(\psi)$, both on the same rollout group:

$$\mathcal{J}_{\mathrm{GRPO}}(\theta, \psi) = \mathcal{J}_{\mathrm{video}}(\theta) + \mathcal{J}_{\mathrm{prompt}}(\psi). \tag{5}$$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and optimization. We evaluate the framework with MiniMax-H3 as the video generator. Phy-Judger is fine-tuned from Qwen3.5-9B for 10 epochs with LoRA rank 32, $\alpha = 64$, and the vision encoder frozen, using AdamW, a learning rate of 2×10^{-5} with cosine decay and 10% warmup, and a batch size of 8. The video policy is fine-tuned for 200 iterations with LoRA rank 32, $\alpha = 64$, AdamW, and a constant learning rate of 3×10^{-4}. The prompt policy uses Qwen3.5-4B with LoRA rank 16 and $\alpha = 32$, a learning rate of 5×10^{-6} decayed linearly to zero, and a GRPO clipping range of 0.01. For joint RL, we use $M = 4$ rewrites per prompt and $N = 4$ videos per rewrite, giving $G = MN = 16$ videos per group, and each batch contains 6 user prompts. Training rollouts use 20 denoising steps with 8 SDE steps, 384×672 resolution, and 124 frames. Phy-MiniMax-H3 denotes MiniMax-H3 after joint RL with Phy-Judger.

Table 1: Preference-data ablation. Each component improves preference accuracy and dimension MAE. Best results in each column are bold.

	Acc. ↑	MAE ↓
zero-shot annotation	59.5	1.42
+ few-shot exemplars	68.0	0.46
+ UNSUPPORTED	71.0	0.43
+ consistency filtering (ours)	**72.5**	**0.42**

Table 2: Phy-Judger ablation on the same data as Table 1. Each component improves both metrics. Best results in each column are bold.

	Acc. ↑	MAE ↓
pointwise, scores only	62.5	0.60
+ rationale	65.0	0.47
+ localization	66.5	0.45
+ pairwise samples (ours)	**72.5**	**0.42**

Reward and evaluation protocol. Phy-Judger provides the physical reward, while HPSv3-general (HPS-G), HPSv3-percentile (HPS-P), and the VideoAlign motion-quality (MQ) and text-alignment (TA) scores track general quality. The physical evaluation uses VBench-2.0 and PhyWorldBench; general metrics check whether physical optimization changes appearance or prompt alignment. VBench-2.0 (Zheng et al., 2025) scores five dimensions: Physics (VB-P), Commonsense (VB-C), Human Fidelity, Controllability, and Creativity; each dimension score is the mean of its sub-dimensions, and Total is the mean of the five dimensions, following the official protocol. PhyWorldBench (Gu et al., 2026) reports semantic adherence (SA), physical commonsense (PC), and their conjunction (Both) under its automatic the Qwen3.8-27B as the evaluator for all models.

Phy-Judger variants are tested on a test set of 200 human-annotated pairs that are not used in training or as few-shot annotation exemplars. Preference accuracy (Acc., %) is the accuracy of the chosen label, and dimension MAE is the mean absolute error between predicted and human dimension scores over applicable dimensions. RL variants are tested on 200 prompts that are not used in RL training. We score the videos with Phy-Judger for physics, and with VideoAlign and HPSv3 for general quality. Phy-Judger scores every video against the original prompt, also for rows with prompt rewriting; because the RL rows optimize this score directly, we also report VB-P and VB-C on the VBench-2.0 prompts as an independent evaluator.

For human evaluation, annotators compare MiniMax-H3 before and after RL on prompts sampled from the two benchmarks, using the same seed for both videos of a pair. Videos are shown side by side in random order without method names; annotators choose the more physically plausible video or a tie. Each of 25 annotators judges 15 pairs (375 judgments). Win, tie, and loss rates are the fractions of pairs in which the video after RL is preferred, the two are tied, or the video before RL is preferred.

5.2 ABLATION STUDIES

We ablate the preference data annotation, training of Phy-Judger, and the joint RL of the generator.

Preference data annotation. Table 1 trains Phy-Judger on data built with different annotation components. The table starts from zero-shot annotation, in which GLM-5.3-Flash annotates every pair without exemplars, must explain every construction label, and no rationale is filtered. It then adds one component at a time: the few-shot exemplars; the UNSUPPORTED option, so that the annotator can reject a construction label; and rationale consistency filtering. Every variant uses 2,000 randomly sampled training pairs, and the zero-shot variant excludes the 200 human-annotated and 400 GPT-6-astra-annotated pairs. All variants train Phy-Judger as in the last row of Table 2. The few-shot exemplars add +8.5 preference accuracy and lower dimension MAE from 1.42 to 0.46, because they show the annotator how humans score each dimension. The UNSUPPORTED option adds +3.0, as it removes pairs whose label is wrong. Consistency filtering adds +1.5, since it removes pairs whose rationale does not support the label.

Phy-Judger learning. Table 2 keeps the training data fixed and changes the supervision of Phy-Judger. Pointwise samples are obtained by splitting each annotated pair into its two videos, each with its part of the rationale and without the preference label. The table starts from training on pointwise samples with dimension scores only, and adds one component at a time: violations with their dimension, severity, and evidence; the time interval and bounding box of each violation; and

Table 3: Joint physical RL ablation on MiniMax-H3 with matched RL prompts and rollout budgets. All components improve physics; SFT reduces general quality. Best results in each column are bold.

	Physics			General quality			
Method	Phy-Judger ↑	VB-P ↑	VB-C ↑	MQ ↑	TA ↑	HPS-G ↑	HPS-P ↑
MiniMax-H3	2.98	75.80	75.23	1.57	2.14	2.94	6.76
+ video-policy RL	3.30	77.03	76.55	1.65	2.18	**3.30**	7.05
+ dimension-wise advantage	3.44	77.20	76.92	1.66	2.16	**3.30**	7.05
+ prompt-policy RL (ours)	**3.62**	**78.83**	**78.04**	**1.81**	**2.73**	3.29	**7.52**
Alternatives							
prompt rewrite	3.12	76.40	75.59	1.61	2.20	3.02	6.99
SFT on good videos	3.34	77.23	76.34	1.30	2.10	2.62	6.42

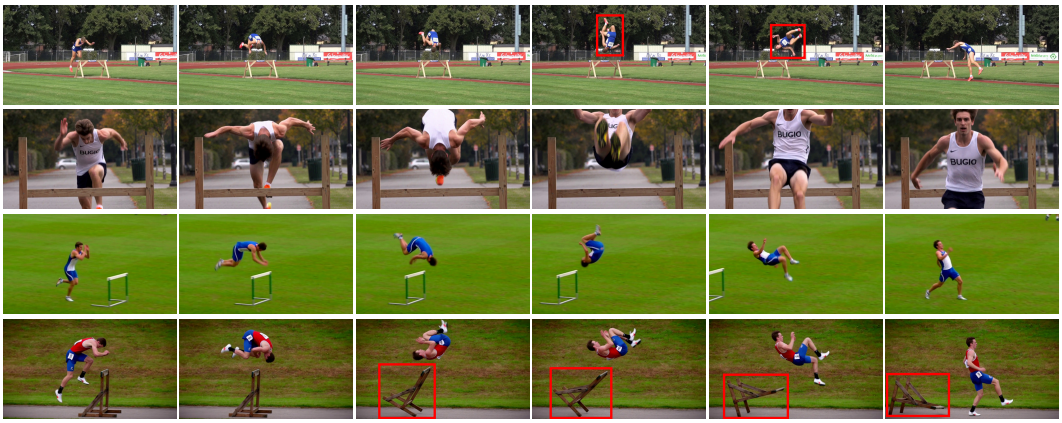


Figure 3: Rows from top to bottom: base model, joint RL, SFT on good videos, and prompt rewrite. The base model changes pose abruptly in mid-air, SFT lowers visual quality, and with prompt rewrite the hurdle still falls over with no force on it.

pairwise samples. The rationale lowers dimension MAE from 0.60 to 0.47, because the judge must name the violations behind each score. Localization adds +1.5 preference accuracy, as the judge must also point to when and where each violation happens. Pairwise samples add +6.0 preference accuracy, since the judge learns to compare two videos directly instead of scoring each one alone.

RL with rewards. All variants in Table 3 use the same training setting. The table starts from the base model MiniMax-H3 and adds one component at a time: video-policy RL with the averaged reward; the dimension-wise advantage in equation 4; and prompt rewriting with a trained prompt policy, which gives joint RL. The prompt rewrite row keeps the prompt policy untrained, which tests whether the gain of joint RL comes from training the prompt policy or from prompt expansion itself. The SFT row replaces online RL with supervised fine-tuning on preferred videos.

Each component raises both the Phy-Judger and the VBench-2.0 scores (Table 3). Video-policy RL adds +0.32 Phy-Judger, since the generator learns directly from Phy-Judger rewards on its own videos. The dimension-wise advantage adds +0.14, because a high score in one dimension no longer hides a violation in another. Prompt-policy RL adds +1.63 VB-P and +0.57 text alignment, as the trained prompt policy spells out the physical conditions that the generator would otherwise miss (Figure 4). The prompt rewrite row gains only +0.06 text alignment: an untrained prompt policy only makes the prompt longer and does not learn which details matter. SFT gains about as much as video-policy RL on physics but lowers MQ by 0.27, because it copies the look of the preferred videos, including real clips (Figure 3).

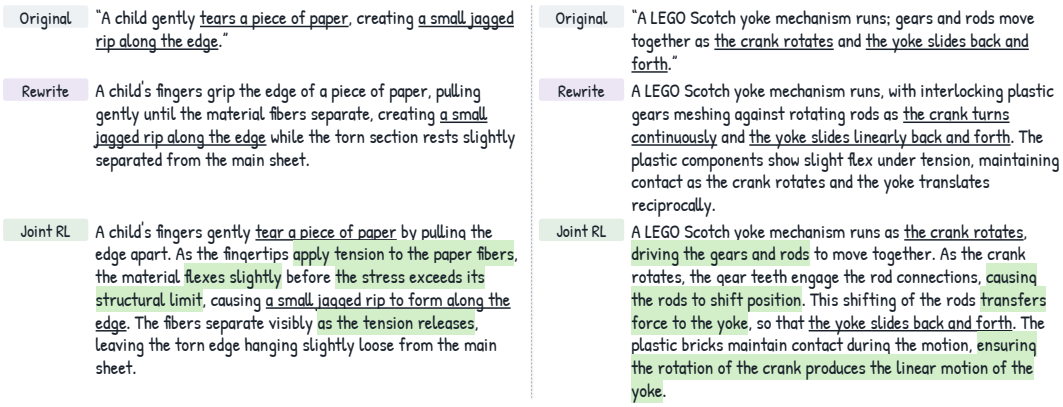


Figure 4: Different prompts strategies. Rewrite is from the untrained prompt policy. The trained prompt policy (Joint RL) keeps the requested event (underlined) and adds the physical conditions that cause it, such as tension, stress, or force transfer (highlighted), which Rewrite does not state.

Table 4: VBench-2.0 results (%). Joint RL raises Physics and Commonsense by 3.03 and 2.81 points, respectively, and Total from 65.67 to 67.17 over MiniMax-H3; other dimensions remain stable or improve. Best results in each column are bold.

Model	Physics ↑	Commonsense ↑	Human Fidelity ↑	Controllability ↑	Creativity ↑	Total ↑
HunyuanVideo	60.20	63.44	82.41	28.60	41.84	55.30
CogVideoX-1.5	63.23	58.18	72.14	29.57	43.66	53.35
Sora	57.19	64.31	87.72	22.09	**60.56**	58.38
Kling 1.6	64.35	65.45	83.56	33.05	48.58	59.00
MiniMax-H3	75.80	75.23	90.13	39.33	47.84	65.67
Phy-MiniMax-H3	**78.83**	**78.04**	**91.72**	**39.50**	47.76	**67.17**

5.3 Main results

Tables 4 and 5 report VBench-2.0 and PhyWorldBench. Figure 5 shows qualitative comparisons, and Table 6 reports the human evaluation.

Joint RL raises VBench-2.0 Physics by +3.03 and Commonsense by +2.81, because Phy-Judger rewards the generator on the same physical dimensions. The other three dimensions change by at most 1.6 points, since Phy-Judger only penalizes physical violations and does not reward any content or style.

On PhyWorldBench, joint RL raises overall PC by +0.048 and Both by +0.043, with gains on both Fundamental and Composite Physics, because the reward penalizes each physical violation that the judge finds. SA also rises by +0.041, as the prompt policy spells out the objects and events of the prompt, while SA is still scored against the original prompt.

Human evaluation. In Table 6, annotators prefer the video after joint RL in 49.1% of comparisons and the base model in 33.1%, with 17.9% ties. This gives a 16.0-percentage-point win–loss margin and an approximately 59.7% win rate among non-tied judgments. The 375 judgments from 25 annotators use matched-seed pairs, randomized presentation order, and hidden method names, providing an assessment independent of the automatic evaluators. This human preference trend agrees with the gains on VBench-2.0 and PhyWorldBench, supporting improvements in visible physical plausibility beyond the learned reward score.

Table 5: PhyWorldBench results. Joint RL raises overall PC/Both by 0.048/0.043 over MiniMax-H3. Best results in each column are bold.

Model	Fundamental Physics			Composite Physics			Anti-Physics			Overall		
	SA ↑	PC ↑	Both ↑	SA ↑	PC ↑	Both ↑	SA ↑	PC ↑	Both ↑	SA ↑	PC ↑	Both ↑
Pika 2.0	**0.603**	0.167	0.149	0.400	0.102	0.095	0.171	**0.029**	**0.029**	0.499	0.133	0.121
Luma	0.444	0.135	0.106	0.305	0.098	0.086	0.095	0.000	0.000	0.368	0.111	0.090
Sora	0.454	0.137	0.102	0.295	0.083	0.067	0.057	0.019	0.010	0.367	0.109	0.082
Kling-1.6	0.432	0.111	0.086	0.273	0.060	0.057	0.086	0.010	0.010	0.350	0.086	0.070
CogVideoX-1.5	0.392	0.137	0.106	0.257	0.057	0.041	0.038	0.000	0.000	0.316	0.099	0.076
Hunyuan	0.397	0.133	0.094	0.292	0.086	0.060	0.057	0.010	0.010	0.331	0.107	0.075
MiniMax-H3	0.564	0.187	0.165	0.413	0.133	0.121	0.162	**0.029**	**0.029**	0.478	0.155	0.138
Phy-MiniMax-H3	**0.603**	**0.235**	**0.206**	**0.460**	**0.197**	**0.181**	**0.190**	**0.029**	**0.029**	**0.519**	**0.203**	**0.181**

Table 6: Human evaluation against MiniMax-H3 (%). Phy-MiniMax-H3 is preferred more often than the base model, with a 16.0-point win–loss margin.

	Physical plausibility		
	Win	Tie	Loss
Phy-MiniMax-H3 vs. MiniMax-H3	49.1	17.9	33.1

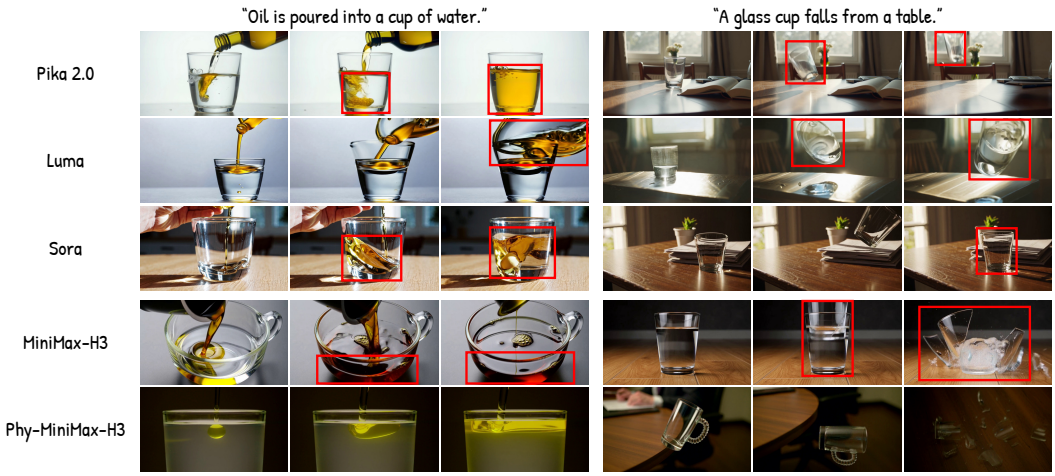


Figure 5: Qualitative comparison on two prompts of PhyWorldBench, three frames each. In the other rows, the oil mixes with or sinks in the water, and the cup floats in the air or breaks on the table. Phy-MiniMax-H3 keeps the oil on top of the water, and the cup falls off the table before it breaks. Red boxes mark the violations.

6 Conclusion

We learn a physical reward from 10,114 preference pairs across 25 sub-scenes, using labels determined by pair construction and rationales that localize visible violations. Phy-Judger generates this evidence before its preference label, making each score grounded in visible violations. The preferences train the reward, which then guides online RL of a video generator and a prompt policy that makes implicit physical conditions explicit. On MiniMax-H3, joint RL improves VBench-2.0 Physics by 3.03 points and Commonsense by 2.81 points, and raises PhyWorldBench overall physical commonsense from 0.155 to 0.203. Human evaluation also favors the resulting videos, supporting the use of learned physical rewards for video post-training.

AI USE STATEMENT

In this work, we used generative AI video tools to generate synthetic data and to clean and reformat data as part of our method: LLMs screen and rewrite prompts, video generators produce the synthetic videos, and VLMs write most rationales (Section 4.1). We have not used generative AI tools for other tasks with required disclosure. Additionally, we used generative AI tools to draft and edit parts of the paper, create figures, and edit code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve the physical plausibility of generated videos. More convincing synthetic videos could also facilitate misinformation or deceptive media. Generated content should therefore be clearly identified as synthetic, and improvements in physical plausibility should not be interpreted as evidence of authenticity or as guarantees of physical correctness for safety-critical applications. Our data construction uses public datasets and generated videos; their use and redistribution should respect source licenses, privacy, and the rights of individuals depicted. The learned reward may inherit biases from source data, pretrained models, and human judgments, so its scores and rationales should be interpreted within the scope of the evaluated scenes rather than as universal assessments of physical validity.

REPRODUCIBILITY STATEMENT

Sections 4 and 5 describe the method and training settings, and Appendix A gives the data construction, prompts, and reward details. We will release the code, checkpoints, and datasets.

REFERENCES

Hritik Bansal, Clark Peng, Yonatan Bitton, et al. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2503.06800.

Omer Bar-Tal, Hila Chefer, Omer Tov, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Aritra Bhowmik, Denis Korzhenkov, Cees G. M. Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. MoAlign: Motion-Centric Representation Alignment for Video Diffusion Models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.19022.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

Yuanhao Cai, Kunpeng Li, Menglin Jia, et al. PhyGDPO: Physics-Aware Groupwise Direct Preference Optimization for Physically Consistent Text-to-Video Generation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2512.24551.

Hongyang Du, Junjie Ye, Xiaoyan Cong, et al. VideoGPA: Distilling Geometry Priors for 3D-Consistent Video Generation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.23286.

Ying Fan, Olivia Watkins, Yuqing Du, et al. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.16381.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, et al. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Jing Gu, Xian Liu, Yu Zeng, et al. PhyWorldBench: A Comprehensive Evaluation of Physical Realism in Text-to-Video Models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.13428.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, et al. T2VPhysBench: A First-Principles Benchmark for Physical Consistency in Text-to-Video Generation. *arXiv preprint arXiv:2505.00337*, 2025. URL https://arxiv.org/abs/2505.00337.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Sihui Ji, Xi Chen, Xin Tao, Pengfei Wan, and Hengshuang Zhao. PhysMaster: Mastering Physical Representation for Video Generation via Reinforcement Learning. *arXiv preprint arXiv:2510.13809*, 2025. URL https://arxiv.org/abs/2510.13809.

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar Geometry Improves Video Generation Models. *arXiv preprint arXiv:2510.21615*, 2025. URL https://arxiv.org/abs/2510.21615.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? Post-Training Newton's Laws with Verifiable Rewards. *arXiv preprint arXiv:2512.00425*, 2025. URL https://arxiv.org/abs/2512.00425.

Haodong Li, Tianfei Ren, Xiaoxiao Ma, et al. VideoCoCo: Code-as-CoT for physically-consistent video generation via an agentic dual-engine system. *arXiv preprint arXiv:2607.27380*, 2026a.

Zizun Li, Haoyu Guo, Runzhe Teng, Chunhua Shen, and Tong He. Geo-Align: Video generation alignment via metric geometry reward. *arXiv preprint arXiv:2605.23903*, 2026b.

Wang Lin, Liyu Jia, Wentao Hu, et al. Reasoning Physical Video Generation with Diffusion Timestep Tokens via Reinforcement Learning. *arXiv preprint arXiv:2504.15932*, 2025. URL https://arxiv.org/abs/2504.15932.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Jie Liu, Gongye Liu, Jiajun Liang, et al. Flow-GRPO: Training Flow Matching Models via Online RL. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2505.05470.

Jie Liu, Gongye Liu, Jiajun Liang, et al. Improving Video Generation with Human Feedback. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2501.13918.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Haoran Lu, Shang Wu, Songling Liu, et al. Phys4D: Fine-Grained Physics-Consistent 4D Modeling from Video Diffusion. *arXiv preprint arXiv:2603.03485*, 2026. URL https://arxiv.org/abs/2603.03485.

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. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.02371.

Wenxu Qian, Chaoyue Wang, Hou Peng, et al. RDPO: Real Data Preference Optimization for Physics Consistency Video Generation. *arXiv preprint arXiv:2506.18655*, 2025. URL https://arxiv.org/abs/2506.18655.

Shuyao Shang, Yuqi Wang, Ruopeng Gao, et al. PhiZero: A World Model Built Around Physical Language. *arXiv preprint arXiv:2607.28624*, 2026. URL https://arxiv.org/abs/2607.28624.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Fu-Yun Wang, Han Zhang, Michael Gharbi, Hongsheng Li, and Taesung Park. PromptRL: Prompt Matters in RL for Flow-Based Image Generation. In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2602.01382.

Haozhe Wang, Cong Wei, Weiming Ren, et al. RationalRewards: Reasoning Rewards Scale Visual Generation Both Training and Test Time. In *Conference on Language Modeling*, 2026b. URL https://arxiv.org/abs/2604.11626.

Jing Wang, Ao Ma, Ke Cao, et al. WISA: World Simulator Assistant for Physics-Aware Text-to-Video Generation. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2503.08153.

Peiyao Wang, Weining Wang, and Qi Li. PhysCorr: Dual-Reward DPO for Physics-Constrained Text-to-Video Generation with Automated Preference Selection. *arXiv preprint arXiv:2511.03997*, 2025b. URL https://arxiv.org/abs/2511.03997.

Weijie Wang, Xiaoxuan He, Youping Gu, et al. World-R1: Reinforcing 3d constraints for text-to-video generation. In *International Conference on Machine Learning*, 2026c.

Wenhao Wang and Yi Yang. VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2403.06098.

Wenhao Wang and Yi Yang. VideoUFO: A Million-Scale User-Focused Dataset for Text-to-Video Generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.01739.

Zeqing Wang, Keze Wang, and Lei Zhang. PhyDetEx: Detecting and explaining the physical plausibility of T2V models. In *European Conference on Computer Vision*, 2026d.

Zhaoqing Wang, Xiaobo Xia, Zhuolin Bie, Jinlin Liu, Dongdong Yu, Jia-Wang Bian, and Changhu Wang. Taming camera-controlled video generation with verifiable geometry reward. In *European Conference on Computer Vision*, 2026e.

Shang Wu, Chenwei Xu, Zhuofan Xia, et al. PhyPrompt: RL-based Prompt Refinement for Physically Plausible Text-to-Video Generation. *arXiv preprint arXiv:2603.03505*, 2026. URL https://arxiv.org/abs/2603.03505.

Zeyue Xue, Jie Wu, Yu Gao, et al. DanceGRPO: Unleashing GRPO on Visual Generation. *arXiv preprint arXiv:2505.07818*, 2025. URL https://arxiv.org/abs/2505.07818.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Qiyuan Zhang, Biao Gong, Shuai Tan, et al. PhysRVG: Physics-Aware Unified Reinforcement Learning for Video Generative Models. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2601.11087.

Dian Zheng, Ziqi Huang, Hongbo Liu, et al. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Kaiwen Zheng, Huayu Chen, Haotian Ye, et al. DiffusionNFT: Online Diffusion Reinforcement with Forward Process. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.16117.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A Comprehensive Benchmark For Physical AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.01989.

A PREFERENCE DATA, PHY-JUDGER, AND RL DETAILS

This section describes the preference data, the training and prompts of Phy-Judger, and the prompts and rewards used in RL.

A.1 PROMPT POOL AND SCENE TAXONOMY

Sources and screening. We pool the text of six public corpora: WISA-80K (Wang et al., 2025a), the VideoPhy-2 training split (Bansal et al., 2026), VideoUFO (Wang & Yang, 2025), OpenVid-1M (Nan et al., 2025), VidProM (Wang & Yang, 2024), and Something-Something V2 (Goyal et al., 2017). After normalization this gives about 146k unique prompts. A text-only LLM screener (Qwen3.8-27B) judges every prompt independently for four physical phenomena: coupled physical interaction, complex human motion, geometric consistency under articulation or viewpoint change, and instance preservation through occlusion. For each phenomenon it returns *yes*, *no*, or *uncertain*, a difficulty level, and verbatim evidence quotes from the prompt. Prompts that describe animation, CGI, gameplay, montage, stop-motion, or time-lapse are excluded. A second, blind call re-checks each positive phenomenon using only that phenomenon's criterion. The screener uses only explicit prompt evidence; it does not rewrite prompts or add events. Near-duplicates are removed by retrieving TF-IDF nearest neighbours and asking the LLM whether two prompts describe the same concrete action process; prompts that differ in action sequence, contact, support, or occlusion are kept.

Scene taxonomy and balanced sampling. Each retained prompt is assigned to one of the 25 sub-scenes in Table 7; prompts that fit no sub-scene are discarded. We then sample prompts so that categories, sub-scenes, and physical phenomena are as balanced as the candidates allow. After that we have 23k prompts. Real–synthetic pairs require a real clip, so their prompts come from the corpora that provide videos (WISA-80K, VideoUFO, and Something-Something V2); the two synthetic–synthetic pair types may use prompts from any corpus.

Table 7: Scene taxonomy of the preference corpus. Each category has five sub-scenes.

Category	Sub-scenes
Human motion	jumping; flips and gymnastics; running and stopping; dance and balance; ball sports
Indoor and everyday objects	dropping and bouncing; stacking and toppling; pouring and splashing; cutting and tearing; rolling and sliding
Industry and robotics	conveyor transfer; grasping and placing; machining and pressing; gears and linkages; lifting and hoisting
Traffic and vehicles	braking; turning; collisions; wheel–road contact; cyclists and pedestrians
Nature and materials	fluid flow; melting and freezing; combustion; animal locomotion; wind, snow, and sand

A.2 VIDEO PAIR CONSTRUCTION

Pair types. The retained corpus has three pair types, and in each the construction rule fixes which video is preferred. Real–synthetic pairs (50%) pair a real clip with a video generated from its realized caption; the real clip is preferred. Strong–weak generator pairs (23%) take two videos for the same prompt from a stronger and a weaker generator, matched in duration and resolution; the stronger generator's video is preferred. Correct–perturbed prompt pairs (27%) use the same generator and seed to render the prompt c and a perturbed prompt $\tilde{c}$ that adds one localized physical violation; the video from c is preferred. A judge could separate real–synthetic pairs by source cues rather than physical quality. Two measures limit this shortcut: the other half of the corpus compares synthetic videos only, and every real–synthetic pair must be supported by a visible violation in the synthetic video; if the teacher finds none, it returns UNSUPPORTED and the pair is removed.

Generators are assigned to a strong or a weak tier before annotation using their published physical-commonsense scores on VideoPhy-2 and PAI-Bench; each strong–weak pair takes one generator

from each tier. All videos are re-encoded with the same codec, frame rate, resolution, metadata policy, and duration range; real clips are trimmed and cropped to the generators' duration range and aspect ratio. Watermarks and source identifiers are removed before annotation and training. Teachers and annotators receive only the original prompt c; the pair type and the perturbed prompt $\tilde{c}$ are withheld. The preferred video is assigned to position A or B uniformly at random, so A and B each account for half of the labels.

Perturbed prompts. For correct–perturbed pairs, an LLM rewrites c into $\tilde{c}$ by adding one visible violation tied to an action or relation already present in the scene. The dimension and error type come from the fixed set in Table 8, and the rewrite follows the prompt below.

Table 8: Error types used to build perturbed prompts.

Dimension	Error types
Mechanics	contact without response; unsupported gravity or support; friction or inertia violation; action at a distance; momentum-transfer violation; buoyancy or density violation; deformation violation; support–motion mismatch
Thermotics	spontaneous phase change; non-conserving phase change; reversed thermal direction
Material behavior	rigidity violation; unsupported fracture; solubility or mixing violation; fluid or particle anomaly
Geometry	non-conserving scale change; depth-order violation; shape or topology corruption; multi-view inconsistency
Motion rationality	perturbed trajectory; speed discontinuity; suppressed action consequence; temporal-progression violation; pose–motion mismatch; rolling-motion mismatch
Instance preservation	instance appearance; instance disappearance; instance merge or split; material non-conservation

Physics-Error Prompt Perturbation

Input: A text-to-video prompt c and the allowed error types.
Task: Identify who or what acts, what is acted on, and which contact, support, motion, or state change the prompt describes. Rewrite c so that exactly one of these existing relations fails in a clearly visible way, using one allowed error type.
Constraints: Keep all entities, the setting, the main activity, the camera, and the style. Add one primary violation only. Do not add entities, tools, forces, heat sources, cuts, or effects to motivate it. Describe visible evidence that rules out the most relevant ordinary explanation. Do not treat foreshortening, flexible deformation, fluid inertia, ordinary deceleration, or brief occlusion as a violation.
Output format: JSON with keys `perturbed_prompt`, `dimension`, `error_type`, `target`, and `added_text`.

A.3 Tiered Rationale Annotation

Rationale schema. A rationale records, for each video, the per-dimension scores and every visible violation. A violation has a dimension, start and end timestamps, a key-frame timestamp inside that interval, a bounding box $[x_1, y_1, x_2, y_2]$ on the key frame in coordinates normalized to $[0, 1000]$, and one sentence of visible evidence. The box covers the entity whose behavior is implausible at the key frame. The same schema is used by human annotators, the model teachers, and Phy-Judger.

Human gold rationales. For each of the 25 sub-scenes, annotators write 16 gold rationales. Annotators see the prompt, both videos at their native frame rate, and the construction label. They mark every violation in the schema above and may reject the label as UNSUPPORTED. The 400 gold rationales define what counts as a violation and how severe it is within each sub-scene.

GPT-6-astra tier. GPT-6-astra annotates 400 further pairs, 16 per sub-scene with the same pair-type mix. For each target pair, it receives the prompt, both time-indexed videos, the construction

label, and 16 human exemplars from the same sub-scene and pair type. Exemplars are given as text (prompt, label, and gold rationale); only the target pair is given as video.

GLM-5.3-Flash tier. GLM-5.3-Flash annotates the remaining pairs with the same prompt. Its 50 exemplars are drawn from the 800 human and GPT-6-astra rationales of the same sub-scene and pair type.

Preference-anchored rationale prompt. GPT-6-astra and GLM-5.3-Flash use the same prompt, following the preference-anchored strategy of RationalRewards (Wang et al., 2026b). For each pair we sample up to four teacher rationales.

Preference-Anchored Rationale Generation

Input: `TEXT`, `VIDEO_A`, `VIDEO_B`, the construction label `Y`, and `EXEMPLARS`: gold rationales from the same sub-scene and pair type, given as text.
Task: Produce the Phy-Judger JSON response that explains `Y` using only visible evidence in `VIDEO_A` and `VIDEO_B`. Follow the exemplars' standard for what counts as a violation and how severe it is, but do not copy their content. Localize every claimed violation with timestamps and a key-frame bounding box, and make all dimension-score orderings consistent with `Y`.
Preference anchor: `Y` is fixed by how the pair was built. Do not change it.
Output format: Return the structured Phy-Judger JSON response. Be concise and avoid generic praise. If no visible violation or difference in the videos supports `Y`, output exactly `UNSUPPORTED`.

Unsupported pairs. A construction label can be wrong, e.g., when both generated videos are equally plausible or equally implausible, or when the synthetic video of a real–synthetic pair shows no visible violation. Such a pair is marked UNSUPPORTED and removed.

Consistency filtering. A rationale is kept only if it passes the schema, interval, and box checks, has no score–decision contradiction, and a checker that sees neither the videos nor the label recovers the construction label from it. A pair whose rationales are all rejected is removed.

Of the 23k pairs from balanced sampling, 54% are UNSUPPORTED and 5% fail consistency filtering. GLM-5.3-Flash finally gives 9,314 valid pairs, and we retain 10,114 pairs in total.

A.4 Phy-Judger Training

We use GPT-6-astra and GLM-5.3-Flash as offline VLM teachers and Qwen3.5-9B as the student. The teachers are used only during data construction; all RL reward calls use the locally deployed 9B student. Phy-Judger is trained on pairwise samples and pointwise samples. Pointwise samples are obtained by splitting each annotated pair into its two videos, each with its part of the rationale and without the preference label. Phy-Judger is tested on 200 human-annotated pairs that are not used in training or as few-shot annotation exemplars. All pairs sharing the same prompt, video, real source clip, near-duplicate clip, or perturbed prompt are assigned to the same split to prevent data leakage.

Each 5–10 second video is sampled uniformly into 20–40 ordered frames. Every frame carries its original frame index and timestamp. To localize violations in time and space, the sampler may additionally supply up to eight dense frames around candidate anomaly peaks; total input is capped at 48 frames per video. Training, validation, reward inference, and all ablations use the same sampling rule.

A.5 Phy-Judger Prompts and Output Scale

Score scale. For each applicable dimension, the judge assigns an integer from 1 to 5:

- 5: fully plausible; no observable violation;
- 4: plausible overall, with only a minor or ambiguous local deviation;
- 3: a noticeable moderate deviation or an incomplete physical outcome;
- 2: a major violation that materially changes the event; and
- 1: a critical, impossible, or event-dominating violation.

The pointwise and pairwise prompts differ only in the input, the decision, and the output keys, so we show them in one box.

Pointwise and Pairwise Judgment

Role: You are Phy-Judger, an evaluator of physical plausibility in generated video.
Input: A text condition and one time-indexed video (pointwise), or two time-indexed videos, A and B (pairwise).
Task: First verify that the described event is observable in every video. Judge only visible evidence under ordinary real-world conditions, unless the text specifies a different environment or an intentionally anti-physical event. Do not reward cinematic style, beauty, or sharpness, except when poor visibility prevents judgment.
Dimensions: *mechanics; thermotics; material behavior; geometry; motion rationality; instance preservation.* Evaluate only applicable dimensions.
Scoring: For every violation, report its dimension, start and end timestamps, a key-frame timestamp inside that interval, a bounding box $[x_1, y_1, x_2, y_2]$ on the key frame normalized to $[0, 1000]$ around the implausible entity, and a brief description of visible evidence. Score each applicable dimension from 1 to 5; use `NA` when it cannot be observed or does not apply. A serious violation must not be offset by unrelated appearance quality.
Decision (pairwise only): Output exactly one of `A`, `B`, or `INVALID`. Choose the video that is more physically plausible; use `INVALID` only when the described event is not observable.
Output format: Return valid JSON only. Pointwise: keys `mode` (`POINTWISE`), `event_present`, `video`, and `confidence`. Pairwise: keys `mode` (`PAIRWISE`), `event_present`, `video_a`, `video_b`, `comparison`, and `confidence`; the `comparison` object contains the decision and concise comparative evidence. Each video object must contain `dimension_scores`, `violations`, `overall_score`, and `concise_evidence`. If the event is not observable, set `event_present` to `false` and `overall_score` to `INVALID`.

A.6 Prompt Collection for Joint Physical RL

We collect a prompt set of 2,000 unique prompts after scene balanced sampling. The prompt set is constructed independently of Phy-Judger to avoid a circular definition in which the judge supplies its own training or RL conditions. RL prompts follow the scene taxonomy in Table 7, with 400 prompts per category spread evenly over its five sub-scenes. Each category is further crossed with the six physics dimensions. All prompts pass the constrained prompt realization in Section A.1.

A.7 RL Reward Weights

For RL, the prediction of Phy-Judger is converted into a continuous reward in $[0, 1]$. In equation 4, Phy-Judger has a total weight of 0.7, split equally over its six dimensions ($w_d = 0.7/6$ each), and the general-quality rewards, HPS-G, HPS-P, MQ, and TA, have a total weight of 0.3.

B More Qualitative Results

Figure 6 shows six more prompts in the same format as Figure 1.



Figure 6: More examples of joint RL with the Phy-Judger reward. For each prompt, the top row is generated before RL, with the violation written and marked in red; the bottom row is generated after joint RL, where the violation is corrected. Tags list the physical dimensions involved.

C LIMITATIONS

We have not studied how our method scales. Phy-Judger is trained on 10k pairs with 400 human rationales, and we do not know how its accuracy changes with more pairs, more human rationales. We have not tested whether the gains hold for other generators, longer training, longer videos, or higher resolution. Phy-Judger writes a rationale before each score, so a reward call is slower than that of a scalar reward model, which makes larger RL runs more costly.