SEEKING PHYSICS IN DIFFUSION NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Do video diffusion models encode signals predictive of physical plausibility? We probe intermediate denoising representations of pretrained Diffusion Transformers (DiTs) and find that physically plausible and implausible videos are partially separable in mid-layer feature space, even at high noise levels. Within-source and perceptual-quality controls suggest that this signal is not fully explained by generator identity or generic visual quality. We distill the signal into a lightweight, backbone-specific physics verifier trained on frozen features and use it in two complementary inference-time mechanisms under a fixed multi-trajectory budget: progressive trajectory selection, which scores trajectories at intermediate checkpoints and prunes weak candidates early, and reward-gradient guidance, which steers surviving trajectories by backpropagating through only the first few DiT blocks. Experiments on PhyGenBench and Physics-IQ across CogVideoX-2B/5B and Wan 2.1-14B show that progressive selection matches verifier-based Best-of-4 on CogVideoX-2B while reducing wall-clock inference time by 37%, whereas reward-gradient guidance substantially improves physical consistency on CogVideoX-5B, all without fine-tuning the video generator.

1 INTRODUCTION

Video diffusion models (Yang et al., 2025b; Kong et al., 2024; Polyak et al., 2024; Brooks et al., 2024) now generate visually realistic videos from text. However, systematic evaluations still reveal frequent violations of basic physical commonsense, including inconsistent gravity, implausible collisions, and unrealistic object dynamics (Kang et al., 2025; Meng et al., 2025; Bansal et al., 2024). This exposes a persistent gap between perceptual realism and physical plausibility.

Existing approaches generally fall into three categories. They either introduce external physical guidance into a frozen generator through physics-conditioned generation (Liu et al., 2024a; Yang et al., 2025a; Yuan et al., 2026c; Gillman et al., 2025; Zhang et al., 2024), modify the generator through physics-aware post-training (Li et al., 2025; Wang et al., 2025b; Zhang et al., 2025a;b; Yuan et al., 2024; Prabhudesai et al., 2024; Wu et al., 2024; Liu et al., 2025c), or perform post-hoc selection by generating N candidates, scoring them with a VLM (Bai et al., 2025; OpenAI, 2024), and retaining the best at a cost that scales linearly with N. Despite their differences, these approaches do not explicitly exploit the generator's intermediate representations as an internal signal of physical plausibility. We revisit this assumption and ask: *does a frozen video diffusion model already encode signals predictive of physical plausibility in its intermediate representations?*

To answer this, we probe intermediate representations of frozen backbones (CogVideoX-2B/5B (Yang et al., 2025b) and Wan 2.1-14B (Team Wan et al., 2025)) along the denoising trajectory, using human-annotated videos generated by diverse models (Bansal et al., 2024). We find that physically plausible and implausible videos are partially separable in mid-layer feature space, even at high noise levels. A lightweight verifier with roughly 1M parameters, trained on these frozen features, achieves an AUC of up to 0.68, and the signal remains after controlling for visual quality and generator identity. These findings indicate that physical plausibility can be assessed before full denoising and pixel-space decoding.

This early and differentiable signal motivates two complementary inference-time mechanisms (cf. Fig. 1). Inspired by inference-time scaling, *progressive trajectory selection* runs N trajectories in parallel and uses the verifier at intermediate checkpoints to retain promising candidates while pruning low-scoring trajectories early, thereby concentrating computation on trajectories more likely

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.

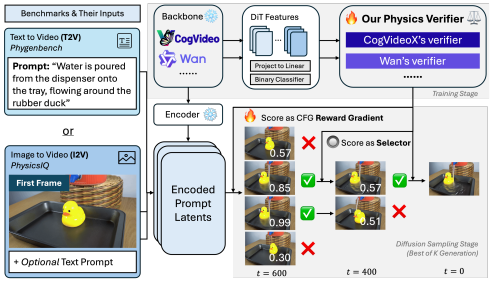


Figure 1: **Overview of our physics-aware inference framework.** For both T2V and I2V generation, a lightweight backbone-specific verifier scores intermediate DiT features. Its predictions are used for reward-gradient guidance and progressive trajectory selection, steering promising trajectories while pruning low-scoring candidates early.

to yield physically plausible videos. Inspired by classifier guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022), *reward-gradient guidance* backpropagates the verifier score to steer each surviving trajectory toward higher predicted physical plausibility. Because the score is computed from an intermediate block, the gradient traverses only the preceding blocks, making guidance substantially cheaper than pixel-space reward guidance (Prabhudesai et al., 2024), which backpropagates through the full transformer and VAE decoder. The two mechanisms are complementary: selection reallocates computation across trajectories, whereas guidance improves each trajectory directly. Both operate on a frozen backbone and require no model fine-tuning.

We evaluate our approach on PhyGenBench (Meng et al., 2025) and Physics-IQ (Motamed et al., 2026) across CogVideoX-2B/5B (Yang et al., 2025b) and Wan 2.1-14B (Team Wan et al., 2025). On CogVideoX-2B, progressive trajectory selection matches Best-of-N physical consistency while reducing wall-clock inference time by 37%. On CogVideoX-5B, where selection alone provides limited gains, reward-gradient guidance substantially improves physical consistency. These results demonstrate that intermediate diffusion features can support both efficient search and direct trajectory steering without fine-tuning the video generator.

Our contributions are:

- We show that physical plausibility is *linearly decodable* from intermediate denoising features, remains detectable within individual generator sources, and is most salient in mid layers under substantial diffusion noise.
- We develop a lightweight physics verifier and two inference-time mechanisms: *progressive trajectory selection*, which prunes low-scoring candidates early, and *reward-gradient guidance*, which directly steers trajectories.
- We validate the framework across multiple video diffusion backbones on physics-oriented T2V and I2V benchmarks, demonstrating improved physical plausibility and more efficient inference without fine-tuning the video generator.

2 Related Work

Physical Understanding in Video Models. Video diffusion models (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023; Bar-Tal et al., 2024; Yang et al., 2025b; Kong et al., 2024; Team Wan et al., 2025) are increasingly framed as world simulators (Brooks et al., 2024; Zhu et al., 2024; Yue et al., 2025; NVIDIA et al., 2025), but strong generation quality does not imply physical competence: Physics-IQ (Motamed et al., 2026) finds generative video models largely fail to predict real-world dynamics, and VideoPhy (Bansal et al., 2024; 2025) and PhyGenBench (Meng et al., 2025) report widespread commonsense violations (see also Kang et al., 2025; Qin et al., 2025), a failure mode that general-purpose suites centered on visual quality and prompt alignment (Huang et al., 2024; Liu et al., 2024b) do not isolate; see Lin et al. (2025) and Liu et al. (2025a) for surveys. They inherit the notion of intuitive physics from developmental psychology (Spelke & Kinzler, 2007; Baillargeon, 2004; Battaglia et al., 2013) and earlier intuitive-physics benchmarks (Bakhtin et al., 2019; Riochet et al., 2022; Bear et al., 2021; Piloto et al., 2022). A complementary line probes

internal representations: intuitive-physics signals emerge in V-JEPA-style self-supervised predictors (Garrido et al., 2025), and diffusion features are known to carry semantic and correspondence information (Baranchuk et al., 2022; Tang et al., 2023; Vélez et al., 2025). For generative video models, LikePhys (Yuan et al., 2026a) reads physics preferences training-free from the denoising likelihood, and concurrent work shows plausibility is linearly decodable from intermediate diffusion features (Esmati et al., 2026; Punzo et al., 2026). These establish that the signal exists but stop short of using it, with Esmati et al. (2026) explicitly leaving probe-guided generation to future work. Closest to our setting, WMReward (Yuan et al., 2026b) steers sampling with an *external* V-JEPA-2 reward. We instead read the generator's *own* intermediate features, so scoring needs no second model, and guidance backpropagates through only the first few DiT blocks.

Physics-Aware Video Generation. One family conditions on explicit physical priors—motion trajectories (Liu et al., 2024a; Yang et al., 2025a; Geng et al., 2025; Zhang et al., 2025c), forces or goals (Gillman et al., 2025; 2026), structured dynamics and kinematic constraints (Yuan et al., 2026c; Wang et al., 2025a; Akkerman et al., 2025; Romero et al., 2025), simulation-driven animation (Xie et al., 2025), or material properties (Zhang et al., 2024); e.g., PhysGen (Liu et al., 2024a) draws trajectories from rigid-body simulation and NewtonGen (Yuan et al., 2026c) from neural Newtonian dynamics. The achievable phenomena are bounded by the expressivity of the chosen prior. A second family updates weights: supervised post-training on curated physics data (Li et al., 2025; Wang et al., 2025b; Zhang et al., 2025a;b), or preference alignment in the spirit of RLHF and DPO (Stiennon et al., 2020; Rafailov et al., 2023; Wallace et al., 2024) from human, VLM, and learned-reward feedback (Yuan et al., 2024; Prabhudesai et al., 2024; Cai et al., 2026; Wu et al., 2024; Liu et al., 2025c). Both demand substantial compute and yield model-specific weights that do not transfer across architectures.

Inference-Time Scaling and Guidance. Test-time search improves quality without touching weights, for language models (Snell et al., 2024) as well as diffusion models, where extra compute is spent searching over noises and trajectories under a verifier (Ma et al., 2025) or on particle-based resampling with intermediate potentials (Wu et al., 2023; Singhal et al., 2025). In video, Video-T1 (Liu et al., 2025b) expands denoising trajectories in a tree, DLBS (Oshima et al., 2025) runs beam search with lookahead rollouts, EvoSearch (He et al., 2025) applies evolutionary search across diffusion and flow generators, and TITAN-Guide (Simon et al., 2025) studies reward-guided inference-time alignment for text-to-video diffusion. A parallel family steers individual trajectories with the gradient of a classifier or reward: classifier guidance trains a noise-aware classifier (Dhariwal & Nichol, 2021), classifier-free guidance makes it implicit (Ho & Salimans, 2022), and training-free variants evaluate an off-the-shelf reward on a one-step estimate of the clean sample (Bansal et al., 2023; Chung et al., 2023; Yu et al., 2023); see Yuan et al. (2023) for a theoretical treatment of reward-directed diffusion. The bottleneck is that reward models correlate poorly with final quality during early denoising—*verifier temporal coherence* in the terminology of Baraldi et al. (2025)—and this is especially acute for physics: evaluators such as PhyGenBench (Meng et al., 2025) and VideoScore (He et al., 2024) are reliable only on decoded videos or require costly rollouts, limiting their use as intermediate rewards.

3 Probing Analysis

Post-hoc selection methods (e.g., Best-of-N) improve physical plausibility by sampling multiple videos and choosing the best, but compute scales linearly with N *full* denoising runs: without a reliable intermediate signal, selection can only happen *after* decoding (Baraldi et al., 2025). We therefore ask when and where physical knowledge emerges along denoising. This tests the premise of our method: if plausibility is decodable from *noisy intermediate* features, trajectories can be scored before decoding. We rule out generator identity (Section 3.2), locate the layers and noise levels that fix the verifier layer and checkpoints (Section 3.3), and control for VAE and quality confounds (Section 3.4).

3.1 Setup

We probe a frozen CogVideoX-2B (Yang et al., 2025b) DiT on VideoPhy (Bansal et al., 2024; 2025), $\sim$4,500 videos from seven text-to-video generators, each annotated for physical commonsense (PC) and semantic accuracy (SA). We use the corresponding binary labels for probing and verifier train-

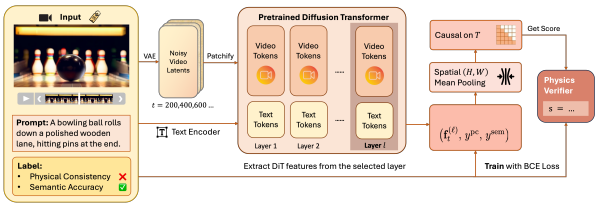
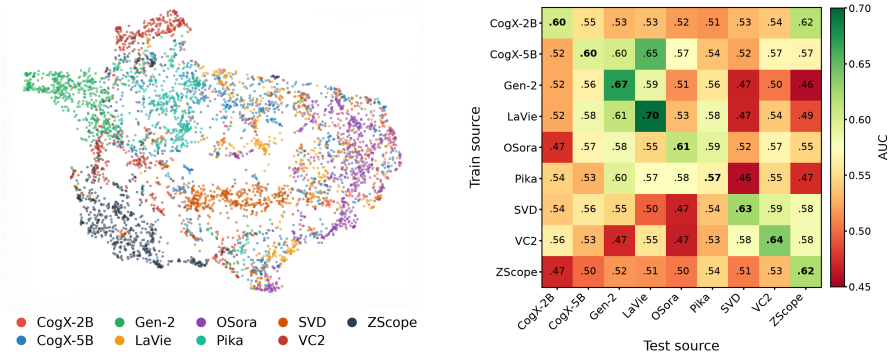


Figure 2: **Frozen-feature extraction and physics-verifier training.** Given a labeled video–prompt pair, we encode the video into VAE latents, add diffusion noise at timestep t, and pass the noisy latents through a frozen diffusion transformer. Hidden states from a selected layer ℓ are extracted, with text tokens removed and video tokens spatially pooled to obtain per-frame features $\mathbf{f}_t^{(\ell)} \in \mathbb{R}^{F\times D}$. A lightweight verifier aggregates these features with causal temporal attention and predicts physical consistency and semantic accuracy, supervised by labels $y^{\text{pc}}, y^{\text{sem}} \in \{0,1\}$ using binary cross-entropy loss. The same frozen features are also used for the probing analysis.



(a) Feature space by generator

(b) Cross-source probing (AUC)

Figure 3: **Source structure in DiT features.** (a) UMAP of CogVideoX-2B DiT features extracted after ℓ=10 at t=200, colored by the original video generator, reveals strong source clustering. (b) Cross-source probing AUC matrix over all nine VideoPhy sources (same features): within-source evaluation (diagonal, 5-fold CV; mean 0.625) is substantially higher than cross-source transfer (off-diagonal mean 0.538).

ing, with PC positives defined as $\text{PC} \geq 3$. As illustrated in Fig. 2, we encode each video into VAE latents (Rombach et al., 2022), add noise at $t \in \{200, 400, 600\}$ (larger t is noisier), run one forward pass, and take hidden states after block $\ell \in \{5, 10, 15, 20, 25\}$; we drop text-conditioning tokens and spatially mean-pool the video tokens into $\mathbf{f}_t^{(\ell)} \in \mathbb{R}^{F\times D}$, with F latent frames and hidden size D. We flatten $\mathbf{f}_t^{(\ell)}$ over time, fit a logistic regression probe for PC, and report mean AUC-ROC under 5-fold cross-validation (CV). Running the *same* probe on raw noised VAE latents isolates the DiT's contribution: outperforming this baseline indicates that physics-related signal becomes more linearly accessible *through* denoising representations.

3.2 Disentangling Source and Physics Signals

Since VideoPhy aggregates seven generators with distinct styles and artifacts, the *video source* is a natural confounder. Figure 3a confirms this: features cluster strongly by generator identity, indicating that source-specific characteristics dominate the representation space, and Figure 3b quantifies it: probes transfer poorly across generators (off-diagonal mean AUC 0.538 vs. 0.625 within source). To control for it, we retrain probes *within* each source (5-fold CV). PC separability survives in every source (AUC 0.534–0.712, Table 1), so DiT features encode PC-relevant information as a secondary signal beneath the dominant source structure. This motivates a *matched-distribution* strategy that we train the verifier only on videos from the generator used at inference.

Table 1: Within-source linear probing AUC using CogVideoX-2B (Yang et al., 2025b) DiT features (ℓ=20, t=200), on the seven non-CogVideoX sources in VideoPhy (Bansal et al., 2024).

Source	# Videos	AUC
LaVie (Wang et al., 2025c)	528	0.712
VideoCrafter2 (Chen et al., 2024)	468	0.631
ZeroScope (Cerspense, 2023)	612	0.613
Gen-2 (RunwayML, 2023)	477	0.612
OpenSora (Zheng et al., 2024)	610	0.597
SVD (Blattmann et al., 2023)	492	0.579
Pika (Pika Labs, 2023)	376	0.534
Overall	3,563	0.652

Table 2: Linear probing AUC (5-fold CV) on CogVideoX-2B DiT features. Every DiT layer improves on VAE probing, by up to +0.101 (ℓ=10, t=600). Best DiT layer per row in **bold**.

Timestep	VAE	ℓ=5	ℓ=10	ℓ=15	ℓ=20	ℓ=25
t=200	0.539	0.569	**0.606**	0.604	0.574	0.579
t=400	0.534	0.545	0.610	0.608	**0.624**	0.594
t=600	0.537	0.565	**0.638**	0.595	0.583	0.585

3.3 Layer and Timestep Analysis

Table 2 reports probing AUC across layers and noise levels. DiT features beat the VAE latent baseline in every configuration (up to +0.101 at ℓ=10, t=600), so the blocks amplify physics-relevant structure. Middle layers carry the most signal (ℓ=10: 0.638), while ℓ=5 and ℓ=25 are weaker, consistent with vision transformers, whose middle layers capture richer semantics (Raghu et al., 2021; Caron et al., 2021). Finally, separability does not degrade with noise: at ℓ=10, AUC at t=600 matches or exceeds t=200, so physics-related structure remains recoverable at moderate noise.

3.4 Control Analyses

A decodable signal is useful for our method only if it comes from the DiT and reflects physics rather than generic visual quality; we test both. First, the signal might originate in the VAE encoder rather than the DiT; probing raw noised latents across all settings shows DiT features consistently win, most starkly on LaVie-sourced videos (0.730 vs. 0.537). Second, physically plausible videos might simply score better under generic evaluators. Using VQAScore (Lin et al., 2024) as a coarse proxy, a probe trained on VQAScore alone reaches only AUC 0.555, and residualizing the linearly VQAScore-correlated component out of the features leaves probing AUC nearly unchanged. The signal is thus enhanced by the DiT and not reducible to perceptual quality.

4 Method

Problem. A frozen video diffusion model maps an initial noise $\mathbf{z}_T$ and a prompt $\mathbf{c}$ to a video $\hat{\mathbf{v}}$. Sampling is stochastic, and physical plausibility can vary substantially across generations from the same prompt, yet the sampler provides no intermediate signal for identifying which trajectory is more likely to obey physical constraints. We ask how to improve the plausibility of the returned video under a fixed initial trajectory budget, while leaving the backbone weights unchanged.

Overview. Our method has three components. (i) A lightweight *physics verifier* predicts physical plausibility from frozen DiT features extracted during denoising (Section 4.1). (ii) At inference time, we denoise N trajectories in parallel and progressively prune low-scoring candidates, concentrating computation on more promising trajectories (Section 4.2). (iii) Between selection checkpoints, we steer each surviving trajectory with the verifier gradient, in the manner of classifier guidance (Section 4.3). Selection *reallocates* computation across trajectories, whereas guidance *modifies* each trajectory.

Notation. Let $\mathbf{z}_t$ denote the noisy latent at diffusion timestep t, $\mathbf{c}$ the text embedding, $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^{S\times D}$ the token-level hidden states after block ℓ, and $\mathbf{f}_t^{(\ell)} \in \mathbb{R}^{F\times D}$ the spatially pooled features defined in Section 3.1. We use $\mathcal{T}$ and $\mathcal{L}$ to denote the candidate diffusion timesteps and DiT blocks, respectively, $\mathcal{C} \subseteq \mathcal{T}$ for selection checkpoints, and $\mathcal{G} \subseteq \mathcal{T}$ for guidance steps.

4.1 PHYSICS VERIFIER

Physical violations unfold over time and propagate to later frames, so we model temporal dependencies with a single causal self-attention layer (Wang et al., 2021): a triangular mask lets frame i attend only to $j \leq i$. This is a temporal inductive bias over *video* frames, so each frame summarizes its past and the last frame summarizes the clip; it is unrelated to diffusion time, and at every denoising step the verifier sees all F latent frames. For each $(t, \ell) \in \mathcal{T} \times \mathcal{L}$ we project $\mathbf{f}_t^{(\ell)}$ to dimension d, add learnable positional embeddings $\mathbf{p}_{1:F}$, and apply causal attention with a residual connection:

$$\tilde{\mathbf{f}}_t^{(\ell)} = \mathrm{Proj}\big(\mathbf{f}_t^{(\ell)}\big) + \mathbf{p}_{1:F} \in \mathbb{R}^{F \times d}, \tag{1}$$

$$\mathbf{y}_t^{(\ell)} = \tilde{\mathbf{f}}_t^{(\ell)} + \mathrm{CausalAttn}\big(\mathrm{LN}(\tilde{\mathbf{f}}_t^{(\ell)})\big). \tag{2}$$

The last-frame representation $(\mathbf{y}_t^{(\ell)})_F$ is therefore a causally aggregated summary of the sequence; a LayerNorm and two-layer MLP map it to a plausibility score $s_t^{(\ell)} = \mathrm{Sigmoid}(\mathrm{MLP}(\mathrm{LN}((\mathbf{y}_t^{(\ell)})_F))) \in [0, 1]$. A second head with identical architecture predicts semantic accuracy.

Training. We train one verifier per layer $\ell \in \mathcal{L}$ on VideoPhy (Bansal et al., 2024), applying forward diffusion at $t \in \mathcal{T}$ and extracting $\mathbf{f}_t^{(\ell)}$. With targets $y^{\text{pc}}, y^{\text{sem}} \in \{0, 1\}$ and head outputs $s_{t,\text{pc}}^{(\ell)}, s_{t,\text{sem}}^{(\ell)}$, the loss is

$$\mathcal{L}^{(\ell)} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbb{E}\big[\lambda_{\text{pc}}\, \mathrm{WBCE}(y^{\text{pc}}, s_{t,\text{pc}}^{(\ell)}) + \lambda_{\text{sem}}\, \mathrm{WBCE}(y^{\text{sem}}, s_{t,\text{sem}}^{(\ell)})\big], \tag{3}$$

where $\mathrm{WBCE}(y, s) = -\, w_y(y \log s + (1-y)\log(1-s))$, w_y corrects class imbalance, and $\lambda_{\text{pc}}, \lambda_{\text{sem}}$ balance the two terms. The frozen backbone means one forward pass per video per timestep suffices; see Section 5.1 for details.

4.2 PROGRESSIVE TRAJECTORY SELECTION

We set ℓ_{best} to the layer in $\mathcal{L}$ whose verifier has the highest validation AUC; this is ℓ=10 on all three backbones, matching the mid-layer peak of Section 3.3 (search details in Section A.2). We then denoise N trajectories from independent noise samples $\{\mathbf{z}_T^{(i)}\}_{i=1}^N$, eliminating implausible candidates along the way. At each checkpoint $t \in \mathcal{C} \subseteq \mathcal{T}$ (ordered high to low noise), the verifier scores all active trajectories from their layer-ℓ_{best} features and retains the top fraction ρ, so the pool shrinks geometrically (Algorithm S1 in Section A.5, Fig. 1). Checkpoints are placed only in the moderate-noise regime where Section 3 found PC separability reliable. We rank on the PC score alone and use the semantic head only as an auxiliary training target, since classifier-free guidance already optimizes prompt alignment.

Unlike pixel-space verifiers, which degrade on noisy intermediate samples (Baraldi et al., 2025), ours is trained on noisy DiT features at the selection timesteps, enabling early pruning where savings are largest.

Computational cost. With T_{steps} DDIM steps, keep ratio ρ, and K roughly evenly spaced checkpoints, the expected number of DiT forward passes is

$$C_{\text{ours}} \approx T_{\text{steps}} \cdot N \cdot \frac{1 - \rho^{K+1}}{(K+1)(1-\rho)}. \tag{4}$$

With N=4, K=2, checkpoints at $t \in \{600, 400\}$, and a 50-step schedule, the pool shrinks $4 \to 2 \to 1$ (20 steps at 4, 10 at 2, 20 at 1), totaling $\approx$120 passes vs. 200 for Best-of-4, an $\approx 1.7\times$ speedup, with negligible verifier overhead.

4.3 REWARD-GRADIENT GUIDANCE

Selection only reallocates compute: each trajectory evolves untouched, so quality is capped by what the N initial noise samples happen to offer. We add a complementary mechanism that steers every

Table 3: **Main results** on PhyGenBench across three backbones. *Official single-seed CogVideoX-2B result from Meng et al. (2025). Pairwise: win rate against the base model among non-tied comparisons; †the judge ties nearly all prompts (157/160, 153/160, 157/160), so these rates are shown only for completeness. Only the final score of Random Sel. was recorded. CogVideoX-2B was the development backbone for selection; guidance and DPO were evaluated on the two larger backbones. Wall-clock costs: Table S5.

Backbone	Method	Final	S1	S2	S3	Mech.	Opti.	Ther.	Mate.	Pairw.
CogVideoX-2B (Yang et al., 2025b)	Base Model	0.370*	—	—	—	0.38	0.43	0.34	0.39	—
	Random Sel.	0.490	—	—	—	—	—	—	—	—
	Best-of-N	0.515	1.96	0.87	**1.73**	0.50	**0.59**	0.46	0.48	—
	+ Selection (Ours)	**0.515**	1.98	**0.91**	1.69	0.49	0.58	0.47	0.49	**66.1%**
CogVideoX-5B (Yang et al., 2025b)	Base Model	0.363	1.54	0.58	1.21	0.283	0.493	0.322	0.308	—
	+ Selection (Ours)	0.365	1.52	0.53	1.30	0.292	0.456	0.256	0.408	60.2%
	+ Sel. + Reward Grad. (Ours)	0.496	**1.79**	**0.95**	**1.71**	**0.417**	**0.607**	**0.456**	**0.467**	**66.7%**
	DPO	0.475	1.74	0.78	1.68	0.375	0.600	0.411	**0.467**	50.3%†
Wan 2.1-14B (Team Wan et al., 2025)	Base Model	0.569	2.05	1.28	1.79	0.525	0.740	0.489	0.458	—
	+ Selection (Ours)	**0.612**	**2.09**	**1.46**	**1.86**	**0.600**	**0.767**	**0.533**	**0.492**	—
	+ Sel. + Reward Grad. (Ours)	0.606	2.09	1.36	1.94	0.575	0.767	0.556	0.475	50.0%†
	DPO	0.558	2.08	1.36	1.53	0.600	0.727	0.433	0.400	50.3%†

active trajectory with the verifier gradient. It starts from the classifier-free guidance (CFG) prediction

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \varnothing) + w_{\text{cfg}}\big(\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \varnothing)\big). \tag{5}$$

CFG ascends the log-likelihood of an *implicit* classifier $p(\mathbf{c} \mid \mathbf{z}_t)$ (Ho & Salimans, 2022). We add the gradient of an *explicit* one, our verifier, exactly as classifier guidance (Dhariwal & Nichol, 2021) adds $\nabla \log p(y \mid \mathbf{z}_t)$ to the score. Two things differ from classic classifier guidance: the classifier is not a separate noise-aware network on latents but a small head on the generator's *own* block-ℓ_{best} features, and the gradient is applied only inside the selection window. Writing the verifier as a differentiable function of the latent, $r_t(\mathbf{z}) = f_\phi(\text{PoolFeat}(\mathbf{h}_t(\mathbf{z}); \ell_{\text{best}})) \in [0, 1]$, we augment Eq. (5) with a matching term:

$$\tilde{\boldsymbol{\epsilon}}_t = \hat{\boldsymbol{\epsilon}}_t - w_{\text{phy}}\, \sigma_t\, \nabla_{\mathbf{z}_t} \log r_t(\mathbf{z}_t), \tag{6}$$

with strength w_{phy} and σ_t the noise standard deviation of $\mathbf{z}_t$ ($\sigma_t = \sqrt{1 - \bar{\alpha}_t}$ for the CogVideoX schedulers). The term is subtracted from the guided model prediction, and the scheduler step is otherwise unchanged. Since $r_t = \text{Sigmoid}(u_t)$, we have $\nabla \log r_t = (1 - r_t)\nabla u_t$: guidance damps automatically once the verifier is confident and is strongest for borderline trajectories, precisely those selection alone cannot rescue.

Cost and schedule. Unlike pixel-space reward guidance, which backpropagates through the full backbone and VAE decoder, our gradient flows only through blocks $1, \ldots, \ell_{\text{best}}$ and the verifier. With ℓ_{best}=10, one guidance step costs $\approx 2\ell_{\text{best}}/L$ extra forward-pass equivalents, a fraction of one denoising step, applied only at $\mathcal{G} \subseteq \mathcal{T}$. Because the verifier is discriminative only at moderate noise (Section 3.3), ascending it elsewhere risks reward hacking and off-manifold drift; we therefore restrict guidance to the selection window (e.g., $t \in [400, 600]$), clip per-frame gradient norms, and keep $w_{\text{phy}} \ll w_{\text{cfg}}$. The two mechanisms compose inside Algorithm S1 (lines 6–7): between checkpoints every trajectory is *improved*, and the checkpoint then *drops* those that failed to respond, so a reward-hacked trajectory that drifts off-manifold is penalized at later, cleaner checkpoints.

5 Experiments

5.1 Setup

Backbones. We evaluate on three frozen video diffusion models: CogVideoX-2B (Yang et al., 2025b) (D=1920, 30 blocks), CogVideoX-5B (Yang et al., 2025b) (D=3072, 42 blocks), and Wan 2.1-14B (Team Wan et al., 2025) (D=5120, 40 blocks). Videos are generated at 480×720 (49 frames) with classifier-free guidance scale 6.0. For trajectory selection we initialize N=4 trajectories and select at $t \in \{600, 400\}$ with keep ratio ρ=0.5, reducing the pool $4 \to 2 \to 1$. Backbone-specific configuration (schedulers, CFG modes, feature hooks) is given in Section A.

Verifiers. Following the matched-distribution principle, we train a separate verifier for each target backbone using backbone-matched videos labeled for physical commonsense (PC) and semantic

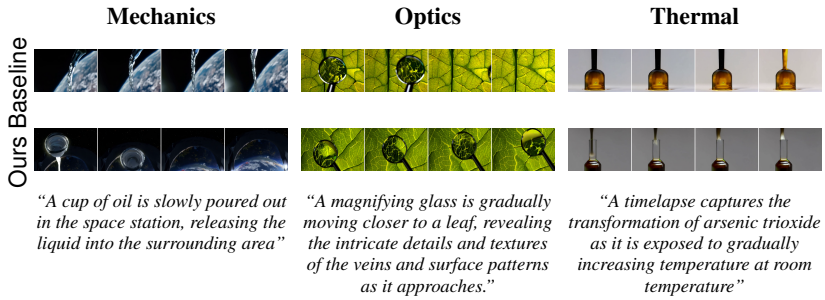


Figure 4: **Qualitative comparison** on PhyGenBench prompts (4 uniformly sampled frames per video). **Mechanics**: baseline pours oil downward under terrestrial gravity; ours forms a floating liquid mass consistent with microgravity. **Optics**: the texture magnified through the baseline's lens appears incoherent with the surrounding leaf venation; ours produces a magnified view whose structure is consistent with the underlying leaf surface. **Thermal and Physical Properties**: baseline shows a liquid stream (incorrect); ours shows rising vapor consistent with sublimation.

accuracy (SA) under the VideoPhy protocols (Bansal et al., 2024; 2025). For each video, we apply forward diffusion at $t \in \{200, 400, 600\}$ and extract frozen features from layer $\ell = 10$. This yields 343, 1,736, and 591 annotated videos for CogVideoX-2B, CogVideoX-5B, and Wan 2.1-14B, respectively, split at the video level into 85% training and 15% validation. The shared architecture has 0.55M–1.37M parameters depending on the backbone width, and each verifier trains in under 10 minutes on a single A100 (hyperparameters in Section A.2).

Baselines. Base Model: base model single-seed generation. **Random Sel.**: identical drop schedule with random verifier scores given. **Best-of-N**: run all N=4 trajectories to completion and select with verifier at t=200. + **Selection (Ours)**: progressive selection (Section 4.2). + **Sel.** + **Reward Grad. (Ours)**: selection plus reward-gradient guidance (Section 4.3). **DPO**: Diffusion-DPO (Wallace et al., 2024) on LoRA adapters, trained on 2K preference pairs built from the same VideoPhy physics annotations. **LikePhys** (Yuan et al., 2026a): training-free likelihood-based physics scorer, compared against our verifier on AUC only.

Benchmarks. PhyGenBench (Meng et al., 2025): 160 text prompts covering 27 physical laws in four categories (mechanics, optics, thermal, material), scored by their PhyGenEval protocol, including VQAScore (Lin et al., 2024) with CLIP-FlanT5-XXL for single-frame phenomena (S1), GPT-4o (OpenAI, 2024) multi-frame verification of event ordering (S2), and GPT-4o naturalness on a 0–3 scale (S3), aggregated per video. **Physics-IQ** (Motamed et al., 2026): conditions on the starting frames of a real scene and scores the continuation against the actual future. **Pairwise**: GPT-4o pairwise judgments against base model, order randomized.

5.2 Main Results

Table 3 reports PhyGenBench results for all backbones and Fig. 4 shows representative examples.

Selection is effective where the verifier is well matched. On CogVideoX-2B, progressive selection reaches 0.515, matching Best-of-N while scoring highest on S2 (0.91), the metric most directly measuring multi-frame physical consistency, and improving over Random Selection (0.490). The gap between Random selection and Base model shows that much of the headline improvement comes from multi-trajectory sampling itself. On Wan 2.1-14B selection gives the largest absolute gain ($0.569 \rightarrow 0.612$, $+7.6\%$), consistent across all four categories and most pronounced on S2 ($+0.18$). On CogVideoX-5B, by contrast, selection is essentially flat ($+0.002$): gains in material properties ($+0.100$) are canceled by losses in optics and thermal. GPT-4o pairwise judgments are nonetheless favorable (60.2% of decided comparisons against the base model, 65/43 with 45 ties over 153 valid pairs), suggesting perceptual improvements the aggregate score does not capture.

Reward gradient helps most where selection cannot. The pattern inverts for gradient guidance. On CogVideoX-5B it lifts the score from 0.363 to 0.496 ($+0.133$), with gains in every category and the largest jump on S2 ($+0.37$). That suggests when all N initial samples are comparably poor and re-ranking cannot help, guidance actively moves each trajectory. On Wan, adding guidance on top

Table 4: **Physics-IQ results** on CogVideoX-5B-I2V (42 of 66 scenarios, three perspectives each; protocol in Section A.6). The gain concentrates in spatial IoU, which scores the regions that move.

Method	Physics-IQ	Spatial IoU	MSE ↓
Baseline	0.242	0.292	0.0064
+ Reward Gradient	**0.262**	**0.317**	**0.0063**

Table 5: **Verifier AUC** on the held-out split of each verifier's training annotations. LikePhys is training-free; ours is trained.

Verifier	CogX-2B	CogX-5B	Wan 2.1-14B
LikePhys	0.554	0.488	0.627
Ours	0.684	0.660	0.657

of selection brings no benefit ($0.612 \rightarrow 0.606$), and the GPT-4o judge ties guidance against the base model on 157 of 160 prompts, i.e. no measurable effect on an already strong backbone. The two mechanisms are thus complementary rather than additive: selection pays off when the candidate pool is diverse (Wan), guidance when all N candidates are comparably poor (CogVideoX-5B).

DPO is not uniformly better. DPO reaches 0.475 on CogVideoX-5B, below reward-gradient guidance (0.496) despite updating model weights. On Wan it *degrades* the base model ($0.569 \rightarrow 0.558$), driven almost entirely by a drop in naturalness (S3: $1.79 \rightarrow 1.53$), a signal of reward over-optimization that inference-time mechanisms avoid, since they never move the model off its own distribution.

Efficiency. Best-of-N runs all four trajectories to completion, whereas selection terminates two at t=600 and one more at t=400. On CogVideoX-2B this is a **37%** wall-clock reduction (490s vs. 778s) at matched output quality (per-component times in Table S5). Checkpoint scoring ($\approx 5\%$ of generation time) is included in these timings. Guidance trades part of this saving for quality, but its truncated backward pass runs at a few steps only.

Where the gains come from. Score-distribution analysis (Section B.2) shows kept and dropped trajectories separate visibly on CogVideoX-5B but overlap substantially on Wan, indicating that on Wan the selection gain is driven more by multi-trajectory diversity than by physics-informed ranking. The Wan case is instead limited by verifier strength, as the Wan verifier is trained on only 591 annotated Wan videos (against 1,736 for CogVideoX-5B), and keeps a held-out AUC of 0.657. However, at that level it is competitive with training-free LikePhys scorer (Table 5).

Zero-shot transfer to I2V. Physics-IQ requires image-to-video generation (CogVideoX-5B-I2V), which shares the T2V DiT architecture. The T2V verifier transfers *zero-shot*: its AUC on I2V features (0.662) matches the T2V value (0.660), and retraining on I2V brings no gain (0.643). With it, selection with reward gradient improves Physics-IQ from 0.242 to 0.262 (Table 4).

6 Discussion and Conclusion

Frozen DiT features carry a measurable signal of physical plausibility that survives within-source and perceptual-quality controls. A verifier on these features lets selection match Best-of-4 on CogVideoX-2B at 37% less wall-clock time, and reward-gradient guidance improves CogVideoX-5B. Together, these results suggest that video diffusion transformers acquire physically relevant structure as a byproduct of learning to denoise.

Several limitations remain. First, the probing signal is moderate, which limits the achievable selection accuracy when candidate trajectories differ only subtly in physical plausibility. Second, the current framework requires training a separate verifier for each target diffusion backbone, limiting plug-and-play transfer across architectures. Third, the verifier is trained on few labeled videos and may not generalize to phenomena underrepresented there. Finally, PhyGenEval relies on GPT-4o (OpenAI, 2024) and inherits the noise of VLM-based judging.

Future work includes simulator-generated training data, larger backbones, and combining our mechanisms with physics-conditioned generation (Zhang et al., 2024; Yuan et al., 2026c; Gillman et al., 2025).

AI USE STATEMENT

In this work, we used generative AI tools for editing and polishing the writing and for assistance with LaTeX formatting and analysis code. We have not used generative AI tools for generating research ideas, for producing or annotating experimental data, or for writing the paper's content; the remaining tasks with required disclosure are not applicable to this work. We have reviewed all AI-assisted work: every edit was checked by the authors, and all code paths that produce reported numbers were verified by rerunning them. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All backbones (CogVideoX-2B/5B, Wan 2.1-14B), benchmarks (PhyGenBench, Physics-IQ), and annotation sources (VideoPhy, VideoPhy-2) are public. The verifier architecture, training hyperparameters, feature-extraction hooks, sampling configurations, seeds, and the exact protocols of the DPO, LikePhys, and VideoREPA baselines are given in Section A, and the Physics-IQ protocol in Section A.6.

REFERENCES

Rick Akkerman, Haiwen Feng, Michael J. Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025.

Renée Baillargeon. Infants' physical world. *Current Directions in Psychological Science*, 2004.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation, 2024.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Lorenzo Baraldi, Davide Bucciarelli, Zifan Zeng, Chongzhe Zhang, Qunli Zhang, Marcella Cornia, Feng Liu, Zheng Hu, and Rita Cucchiara. Verifier matters: Enhancing inference-time scaling for video diffusion models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2025.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 2013.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Fei-Fei Li, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators.
urlhttps://openai.com/research/video-generation-models-as-world-simulators, 2024.

Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, Xiaoliang Dai, Xuan Ju, Alan Yuille, and Ji Hou. PhyGDPO: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Cerspense. zeroscope_v2: A watermark-free modelscope-based video model. `https://huggingface.co/cerspense/zeroscope_v2_576w`, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show. *arXiv preprint arXiv:2606.05328*, 2026.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Deqing Sun, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, and Andrew Owens. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.19386.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, Charles Herrmann, and Chen Sun. Goal force: Teaching video models to accomplish physics-conditioned goals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2601.05848.

Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling image and video generation via test-time evolutionary search. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.17618.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning (ICML)*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.09595.

Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, and Donglin Wang. Exploring the evolution of physics cognition in video generation: A survey, 2025.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024.

Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, Ajmal Mian, Mubarak Shah, and Chang Xu. Generative physical AI in vision: A survey, 2025a.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation. In *International Conference on Computer Vision (ICCV)*, 2025b. arXiv:2503.18942.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems*, 2025c. arXiv:2501.13918.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024a. arXiv:2409.18964.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning (ICML)*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958, 2026.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI, 2025.

OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card, 2024.

Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2501.19252.

Pika Labs. Pika, 2023. `https://pika.art`.

Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6: 1257–1267, 2022.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2024.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Samuele Punzo, Niccolò Caselli, Ippokratis Pantelidis, Francesco Massafra, Salvatore Lo Sardo, and Mohammadreza Salehi. Do video foundation models understand intuitive physics? a layerwise probing analysis. *arXiv preprint arXiv:2606.09646*, 2026.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, and Ruimao Zhang. WorldSimBench: Towards video generation models as world simulators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50338–50362. PMLR, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *NeurIPS*, 2021.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

David Romero, Ariana Bermudez, Viacheslav Iablochnikov, Hao Li, Fabio Pizzati, and Ivan Laptev. Learning to generate rigid body interactions with video diffusion models. *arXiv preprint arXiv:2510.02284*, 2025.

RunwayML. Runway gen-2: Text to Video Generation. `https://runwayml.com/research/gen-2`, 2023. Accessed: 2026-03-04.

Christian Simon, Masato Ishii, Akio Hayakawa, Zhi Zhong, Shusuke Takahashi, Takashi Shibuya, and Yuki Mitsufuji. Titan-guide: Taming inference-time alignment for guided text-to-video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *International Conference on Machine Learning (ICML)*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters, 2024.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 2007.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Pedro Vélez, Luisa F. Polanía, Yi Yang, Chuhan Zhang, Rishabh Kabra, Anurag Arnab, and Mehdi S. M. Sajjadi. From image to video: An empirical study of diffusion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative physics for controllable and physics-grounded video generation. In *Advances in Neural Information Processing Systems*, 2025a. arXiv:2509.20358.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LaVie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133: 3059–3078, 2025c.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Boosting text-to-video generative model with mllms feedback. In *Advances in Neural Information Processing Systems*, 2024.

Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative cartoon animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2501.16550.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. VLIPP: Towards physically plausible video generation with vision and language informed physical prior. In *International Conference on Computer Vision (ICCV)*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025b.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. InstructVideo: Instructing video diffusion models with human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2312.12490.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. In *International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2510.11512.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026b.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. In *International Conference on Learning Representations (ICLR)*, 2026c. arXiv:2509.21309.

Jingtong Yue, Ziqi Huang, Zhaoxi Chen, Xintao Wang, Pengfei Wan, and Ziwei Liu. Simulating the visual world with artificial intelligence: A roadmap, 2025.

Ke Zhang, Cihan Xiao, Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Think before you diffuse: Infusing physical rules into video diffusion. *arXiv preprint arXiv:2505.21653*, 2025a.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3D objects via video generation. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2404.13026.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. In *Advances in Neural Information Processing Systems*, 2025b.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. Is Sora a world simulator? A comprehensive survey on general world models and beyond, 2024.

APPENDIX

This appendix provides additional implementation details (cf. Section A), additional experimental analysis (cf. Section B), qualitative examples (cf. Section C), and failure cases (cf. Section D) that complement the main paper.

A ADDITIONAL IMPLEMENTATION DETAILS

This section provides additional details on the physics verifier architecture, feature extraction pipeline, and video generation setup, complementing the method and experiment sections of the main paper. As reported in the main paper, beyond the primary experiments on CogVideoX-2B (Yang et al., 2025b) we further evaluate our framework on CogVideoX-5B (Yang et al., 2025b) and Wan 2.1-14B (Team Wan et al., 2025). The physics verifier architecture is shared across all three settings, with only the input projection layer adapted to the hidden dimension of each backbone.

A.1 PHYSICS VERIFIER ARCHITECTURE

The physics verifier takes as input the spatially pooled per-frame DiT features extracted from layer ℓ at denoising timestep t, denoted by $\mathbf{f}_t^{(\ell)} \in \mathbb{R}^{F\times D}$, where $F = 13$ is the number of latent frames and D is the hidden dimension of the backbone. Table S1 presents the full layer-by-layer architecture, while Table S2 summarizes the backbone-specific configurations.

Table S1: Architecture of the proposed physics verifier. The input projection layer is adapted to the hidden dimension D of the corresponding backbone, while all remaining components are shared across backbones.

Component	**Operation**	**Output Shape**
Input	Spatially pooled per-frame DiT features	$[B, F, D]$
Input Projection	$\mathrm{Linear}(D \to d)$	$[B, F, d]$
Positional Embedding	Add learnable $\mathbf{p}_{1:F} \in \mathbb{R}^{F\times d}$	$[B, F, d]$
Layer Normalization	Pre-normalization	$[B, F, d]$
Causal Self-Attention	Multi-head ($h = 4$) with a causal mask	$[B, F, d]$
Residual Connection	$\tilde{\mathbf{f}} + \mathrm{CausalAttn}(\mathrm{LN}(\tilde{\mathbf{f}}))$	$[B, F, d]$
Sequence Pooling	Select the final-frame representation	$[B, d]$
Layer Normalization	Apply LayerNorm	$[B, d]$
Classifier Head	$\mathrm{Lin}(d{\to}128){\to}\mathrm{GELU}{\to}\mathrm{Lin}(128{\to}1)$	$[B, 1]$
Output	Sigmoid physical plausibility score	$[B, 1]$

Table S2: Backbone-specific configurations of the three DiT backbones evaluated in this work. The physics verifier architecture ($d = 256$, $h = 4$) is shared across all settings; only the input projection layer and the feature extraction hook differ across backbones.

	CogVideoX-2B	**CogVideoX-5B**	**Wan 2.1-14B**
Hidden dimension D	1920	3072	5120
Number of transformer layers	30	42	40
Text tokens in sequence	Yes (226 tokens)	Yes (226 tokens)	No (cross-attention)
Feature extraction hook	`transformer_blocks`	`transformer_blocks`	`blocks`
Positional encoding	Learned	Rotary (RoPE)	—
Sampling scheduler	DDIM	DDIM	UniPCMultistep
Verifier parameters	~ 0.55M	~ 0.85M	~ 1.37M
Verifier input shape	$[13, 1920]$	$[13, 3072]$	$[13, 5120]$

Design choices. The causal self-attention mask ensures that the representation at frame i attends only to frames $j \leq i$, thereby enforcing temporal causality. This design is motivated by the observation that physical violations often emerge as progressively accumulated inconsistencies over

Table S3: Training hyperparameters of the physics verifier, shared across all backbones.

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	0.01
Batch size	32
Learning-rate schedule	OneCycleLR (10% warmup + cosine annealing)
Early stopping	Patience of 20 epochs on validation AUC
Loss function	Weighted BCE ($\mathrm{pos_weight} = n_{\mathrm{neg}}/n_{\mathrm{pos}}$)
Training timesteps	$\{200, 400, 600\}$ (pooled)
Projection dimension d	256
Number of attention heads	4

time (e.g., an object gradually deviating from a physically plausible trajectory across successive frames). Using the final-frame representation as the sequence summary allows the verifier to aggregate information from all preceding frames while preserving causal ordering. We use $d = 256$ and $h = 4$ attention heads in all experiments. The parameter count is dominated by the input projection layer, $\mathrm{Linear}(D \to d)$, and therefore scales with the backbone hidden dimension D: from approximately 0.55M parameters for CogVideoX-2B ($D = 1920$) to approximately 1.37M for Wan 2.1-14B ($D = 5120$). In every setting, the verifier adds at most $\sim 0.03\%$ additional parameters relative to its backbone.

Backbone-specific differences in feature extraction. The three backbones differ primarily in their text-conditioning mechanisms. CogVideoX-2B and CogVideoX-5B prepend 226 T5 text tokens to the video-token sequence, and we remove these text tokens during feature extraction. By contrast, Wan 2.1-14B uses cross-attention for text conditioning, so its hidden states contain only video tokens. In addition, CogVideoX-5B requires explicit rotary positional embeddings (`image_rotary_emb`) at each transformer forward pass; otherwise, generation collapses to all-black outputs. Finally, because the UniPCMultistep scheduler used by Wan 2.1-14B is stateful, we maintain a separate deep-copied scheduler for each parallel trajectory during multi-trajectory generation.

A.2 Training Details

Training data. Following the matched-distribution principle, each verifier is trained only on videos generated by its own backbone: 343 CogVideoX-2B generations with binary physical-commonsense labels from VideoPhy (Bansal et al., 2024), 1,736 CogVideoX-5B generations, and 591 Wan 2.1 generations with 5-point PC ratings from VideoPhy-2 (Bansal et al., 2025), binarized at PC$\geq$3. Each pool is split 85%/15% into train/validation (seed 42). For every video we extract features with the corresponding frozen backbone at $t \in \{200, 400, 600\}$, tripling the number of training samples; features are stored as `fp16` tensors of shape $[13, D]$.

Training hyperparameters. Table S3 summarizes the training configuration, which is kept identical across all three backbones. To address class imbalance (36.9% positive for CogVideoX-2B, 66.4% for CogVideoX-5B), we use weighted binary cross-entropy with $\texttt{pos_weight} = n_{\mathrm{neg}}/n_{\mathrm{pos}}$. Training completes in under 10 minutes on a single GPU for each backbone.

Layer selection. We train separate physics verifiers for candidate layers ($\ell \in \{5, 10, 15, 20, 25\}$ for CogVideoX-2B; up to $\ell = 35$ for the deeper CogVideoX-5B and Wan backbones) and pick the layer with the highest validation AUC. The causal verifier consistently outperforms the flat linear probe of the main paper, which we attribute to its causal temporal modeling. The selected layers and resulting AUCs are the ones reported in the main paper's verifier comparison table.

Reward-gradient guidance. Guidance uses the same verifier as selection. At each guidance step we recompute the truncated forward pass through blocks $1, \ldots, \ell_{\mathrm{best}}$ with gradients enabled, evaluate $\log r_t$, and backpropagate to the latent; the resulting gradient is normalized by a per-frame

norm clip of 1.0 and injected with strength $w_{\text{phy}} = 0.5$ inside the window $t \in [400, 600]$, i.e. only between the two selection checkpoints. Everything outside the truncated subgraph stays under `torch.no_grad`.

DPO baseline. The Diffusion-DPO baseline trains LoRA adapters (rank 64, α 64) on the attention projections of the frozen backbone with the standard Diffusion-DPO objective ($\beta = 2500$), using 2,000 same-caption preference pairs built from the VideoPhy physics annotations, AdamW at learning rate 10^{-5}, gradient accumulation 2, and 1,000 optimizer steps. The frozen reference model is obtained by disabling the adapters, so no second copy of the backbone is required.

LikePhys protocol. The training-free likelihood baseline scores a video by the negative denoising error $-\mathbb{E}_t\|\hat{v} - v^*\|^2$ at $t \in \{200, 400, 600\}$ under caption conditioning, with noise draws shared across videos (paired comparison) and four draws per timestep. The noising and target follow each backbone's native parameterization (v-prediction for CogVideoX, flow matching for Wan). AUC is computed against the same annotations as the trained verifiers.

A.3 Feature Extraction Pipeline

DiT feature dimensions. CogVideoX (both 2B and 5B) processes video latents as a flattened sequence of spatio-temporal tokens. For 49-frame videos at 480×720 resolution, the VAE produces latent tensors of shape $[13, C, 60, 90]$, corresponding to 13 temporal frames with spatial resolution 60×90. The DiT then patchifies and flattens these latents into $13 \times 30 \times 45 = 17{,}550$ video tokens, with 226 text-conditioning tokens prepended to the sequence. At each layer, we extract the full hidden state $\mathbf{h}^{(\ell)} \in \mathbb{R}^{(226+17{,}550)\times D}$, discard the first 226 text tokens, reshape the remaining video tokens into $[13, 30 \times 45, D]$, and mean-pool over the spatial dimension to obtain $\mathbf{f}^{(\ell)} \in \mathbb{R}^{13\times D}$. For Wan 2.1-14B, text conditioning is applied exclusively through cross-attention, so the hidden states contain only video tokens. The extraction pipeline is therefore simpler: we reshape the video-token sequence from $[B, T \cdot H \cdot W, D]$ to $[B, T, H \cdot W, D]$ and then mean-pool over the spatial dimension. No text-token removal is required.

Hook mechanism. Features are extracted via a PyTorch forward hook registered on the target transformer block. For CogVideoX, the hook is attached to `transformer.transformer_blocks[`ℓ`]`, whereas for Wan it is attached to `transformer.blocks[`ℓ`]` due to differences in module naming. The hook records the *output* of the transformer block after the residual connection, rather than intermediate attention or MLP activations. Because these features are already computed during the standard forward pass, the additional computational overhead is negligible.

Text conditioning. CogVideoX adopts a hybrid text-conditioning scheme. Specifically, text information is injected via cross-attention in every transformer block, while text tokens are also concatenated with video tokens in the input sequence. During feature extraction, we provide the text prompt and therefore obtain text-conditioned features, but discard the explicit text tokens from the extracted hidden states, retaining only the video-token representations. As a result, textual information is reflected implicitly through the conditioned video features rather than through explicit text-token features. By contrast, Wan 2.1-14B performs text conditioning exclusively through cross-attention, and text tokens are never concatenated with the video-token sequence. This removes the need for text-token stripping and leads to a simpler feature extraction pipeline.

A.4 Video Generation Configuration

Backbones. All backbones generate 49-frame videos at 480×720 with classifier-free guidance scale 6.0. The CogVideoX backbones use the DDIM scheduler and Wan 2.1-14B uses UniPCMultistep.

Trajectory selection. For progressive trajectory selection, we initialize $N = 4$ parallel trajectories using independent random seeds (defined by the base seed, prompt index, and trajectory index), and perform selection at checkpoints $t \in \{600, 400\}$ with a keep ratio of $\rho = 0.5$. As a result, the number of active trajectories decreases from $4 \to 2 \to 1$ across the two selection stages. This configuration is used consistently across all three backbones. Checkpoints are not placed at t=800

Table S4: Summary of the video generation configuration for the three backbones evaluated in this work.

Parameter	CogVideoX-2B	CogVideoX-5B	Wan 2.1-14B
Resolution	480×720	480×720	480×720
Number of frames	49 (13 latent)	49 (13 latent)	49 (13 latent)
Sampling scheduler	DDIM	DDIM	UniPCMultistep
Guidance scale	6.0	6.0	6.0
Number of trajectories (N)	4	4	4
Selection checkpoints ($\mathcal{C}$)	$\{600, 400\}$	$\{600, 400\}$	$\{600, 400\}$
Keep ratio (ρ)	0.5	0.5	0.5
Verifier layer	10	10	10
Primary hardware	A5000 (24 GB)	L40S (48 GB)	H200 (80 GB)

because discrimination at high noise levels is unreliable, nor at t=200 because late selection in the schedule yields minimal computational savings.

Scoring at checkpoints. At each checkpoint, we perform one additional forward pass for each active trajectory using only the conditional (positive) prompt embedding, without classifier-free guidance, and the physics verifier scores the captured layer-ℓ features. With N=4 and two checkpoints this amounts to six extra passes on top of the ≈120-pass selection schedule (≈5%), already included in all reported wall-clock times; the verifier head itself adds less than 0.01 s per trajectory.

VAE decoding. Only the final surviving trajectory is decoded by the VAE. We enable VAE tiling but disable slicing, as we observed that enabling both simultaneously can produce solid-color artifacts in the 3D causal VAEs used by both CogVideoX and Wan.

CogVideoX-5B specific notes. CogVideoX-5B requires explicit rotary positional embeddings (`image_rotary_emb`) computed using a 2D grid that matches the spatial latent dimensions. These embeddings must be provided at every transformer forward pass; omitting them leads to all-black outputs. By contrast, CogVideoX-2B uses learned positional embeddings and does not require this additional input.

Wan 2.1-14B specific notes. Wan 2.1-14B (approximately 28 GB in bf16) fits on a single 80 GB GPU (H200). We observe that distributing Wan across multiple devices can introduce tiling and blurring artifacts.

Reproducibility. For prompt i, the first trajectory is initialized with seed $42 + i$, and subsequent trajectories use seeds $42 + i + 1$, $42 + i + 2$, and so on. This setup ensures that trajectory 0 in our method shares the same initial noise as the single-seed baseline, enabling a controlled comparison of the effect of trajectory selection.

Hardware and timing. Table S5 provides wall-clock time breakdowns per video across all three backbones, measured from SLURM job logs. For CogVideoX-5B, the per-step cost is ∼4.6s per trajectory on L40S. Under the 4→2→1 schedule, sampling splits into a four-trajectory phase (∼220s), a two-trajectory phase (∼55s), and a single-trajectory phase (∼55s), plus six scoring forwards (∼27s). For Wan 2.1-14B on H200, the per-step cost is ∼3.9s per trajectory despite 5.6× more parameters than CogVideoX-2B, because batch-1 inference is memory-bandwidth bound and H200 provides ∼5.5× higher bandwidth (4,800 GB/s HBM3e vs. 864 GB/s on L40S). The two selection checkpoints reduce per-step cost from ∼15.5s (4 trajectories) to ∼8s (2) to ∼5.5s (1).

A.5 Progressive Selection Algorithm

Algorithm S1 gives the full procedure. Lines 6–7 apply reward-gradient guidance at steps $t \in \mathcal{G}$ (the window between the two checkpoints); with w_{phy}=0 it reduces to selection alone.

Table S5: Wall-clock time breakdown per video across backbones and modes.

Backbone (GPU)	Component	Baseline	Rand. Sel.	Best-of-4	Ours
CogVideoX-2B (A5000 24 GB)	Sampling	183s	454s	732s	448s
	Scoring	—	—	3s	3s
	Decode + offload	60s	60s	60s	60s
	Total	**204s**	**477s**	**778s**	**490s**
CogVideoX-5B (L40S 48 GB)	Sampling	137s	—	—	357s
	Decode + offload	12s	—	—	12s
	Total	**149s**	—	—	**374s**
Wan 2.1-14B (H200 80 GB)	Sampling	117s†	—	—	391s
	Decode	1s	—	—	1s
	Total	**118s**†	—	—	**393s**

†Estimated on H200. Measured baseline on 2×L40S is ∼305s due to cross-GPU overhead.

Algorithm S1 Progressive Trajectory Selection with Reward-Gradient Guidance

Require: Text embedding $\mathbf{c}$; number of trajectories N; checkpoints $\mathcal{C}$; keep ratio ρ; selected layer ℓ_{best}; physical verifier f_ϕ; guidance steps $\mathcal{G}$ and strength w_{phy}; maximum diffusion timestep T
Ensure: Generated video $\hat{\mathbf{v}}$

Sample initial noises $\mathbf{z}_T^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $i = 1, \dots, N$
Initialize active trajectory set $\mathcal{A} \leftarrow \{1, \dots, N\}$
for $t = T, T-1, \dots, 1$ **do**
  for $i \in \mathcal{A}$ **do**
    $\hat{\boldsymbol{\epsilon}}_t^{(i)}, \mathbf{h}_t^{(i)} \leftarrow \text{DiTForward}(\mathbf{z}_t^{(i)}, t, \mathbf{c})$ ▷ CFG, Eq. (5)
    if $t \in \mathcal{G}$ **then** ▷ reward-gradient guidance
      $\hat{\boldsymbol{\epsilon}}_t^{(i)} \leftarrow \hat{\boldsymbol{\epsilon}}_t^{(i)} - w_{\text{phy}}\, \sigma_t\, \text{clip}\big(\nabla_{\mathbf{z}_t} \log r_t(\mathbf{z}_t^{(i)})\big)$ ▷ Eq. (6)
    end if
    $\mathbf{z}_{t-1}^{(i)} \leftarrow \text{SchedulerStep}(\mathbf{z}_t^{(i)}, \hat{\boldsymbol{\epsilon}}_t^{(i)}, t)$
    $\mathbf{f}_t^{(i)} \leftarrow \text{PoolFeat}(\mathbf{h}_t^{(i)}; \ell_{\text{best}})$
  end for
  if $t \in \mathcal{C}$ **then**
    for $i \in \mathcal{A}$ **do**
      $s_t^{(i)} \leftarrow f_\phi(\mathbf{f}_t^{(i)})$
    end for
    Keep only the top $\lceil \rho |\mathcal{A}| \rceil$ trajectories ranked by $s_t^{(i)}$
  end if
end for
Let i^* be the remaining trajectory in $\mathcal{A}$
return $\hat{\mathbf{v}} \leftarrow \text{VaeDecode}(\mathbf{z}_0^{(i^*)})$

A.6 Physics-IQ Evaluation Protocol

Physics-IQ conditions the model on the switch frame of a real recorded scene and scores the generated continuation against the actual future. We generate with CogVideoX-5B-I2V, conditioning on the official switch frames and scene descriptions, and evaluate with the official toolbox (binary masks, spatial/spatiotemporal IoU, weighted spatial IoU, MSE). Generated clips are resampled to the benchmark's 30 FPS, 5-second format before scoring. We cover 42 of the 66 scenarios, each from all three camera perspectives (252 videos per method); the remaining scenarios were not generated for both methods under our compute budget, and the evaluator's completeness checks are relaxed accordingly, scoring only complete scenario/perspective sets.

Method	Final	S1	S2	S3	Time
VideoREPA	0.492	**1.87**	**0.97**	1.63	1×
DPO (LoRA)	0.475	1.74	0.78	1.68	1×
+ Selection + Reward Gradient (Ours)	**0.496**	1.79	0.95	**1.71**	2.0×

Table S6: **VideoREPA comparison** on PhyGenBench (CogVideoX-5B, all methods measured by us under one shared protocol, seed 42, and scored with the same PhyGenEval pipeline). Time is relative single-prompt inference cost.

B Additional Experimental Analysis

B.1 Comparison with VideoREPA

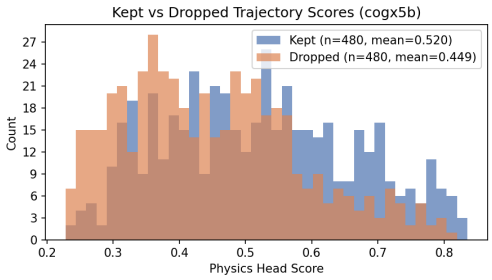
VideoREPA (Zhang et al., 2025b) is a training-based alternative that fine-tunes CogVideoX-5B by aligning token-relation structure in DiT features with a self-supervised video encoder (VideoMAEv2), and thus complements the training-based DPO baseline of the main paper: both spend training compute to move the backbone's weights, whereas our mechanisms spend inference compute on a frozen backbone. We evaluate the officially released VideoREPA-5B adapter with its official inference configuration (LoRA fused at scale $\alpha/r = 64/128$), generating all 160 PhyGenBench prompts under one shared protocol (seed 42, single trajectory) and scoring with the same PhyGenEval pipeline.

VideoREPA is the strongest training-based reference we measure (0.492), essentially matching our inference-time result (0.496): it leads on single-frame semantics and event ordering but trails on naturalness (S3 1.63 vs. our 1.71). The two are also not exclusive, since our mechanisms operate on frozen weights and could be applied on top of VideoREPA's fine-tuned backbone; we leave this combination to future work. Notably, the two prompts on which our base model collapses to degenerate near-constant videos also collapse under VideoREPA's fine-tuned weights, suggesting these failure modes are rooted in the base model rather than addressable by light physics-oriented fine-tuning.

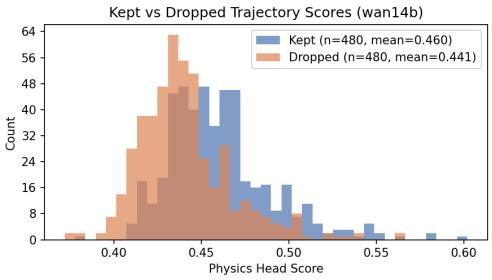
B.2 Score Distribution Analysis

Figure S1 compares the verifier's scoring behavior on CogVideoX-5B and Wan 2.1-14B, complementing the cross-backbone results reported in the main paper. On 5B, kept and dropped trajectories show visible separation (Δmean$=0.071$) and per-prompt score spread averages 0.125, indicating the head provides a meaningful selection signal. On Wan, the two distributions nearly completely overlap (Δmean$=0.019$) and score spread averages only 0.034 ($3.7\times$ smaller), confirming that the Wan verifier operates near chance and the PhyGenBench improvement (+0.043) is attributable to multi-trajectory diversity rather than physics-informed guidance.

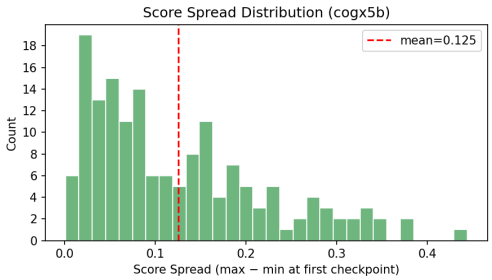
Per-category analysis. Figure S2 breaks down verifier scores by the 27 physical law categories in PhyGenBench. On CogVideoX-5B, the verifier exhibits meaningful category-level specialization. It achieves the highest confidence on phenomena characterized by salient temporal state changes: Color mixing (0.693), Deposition (0.675), Boiling (0.668), Tyndall Effect (0.635), and Reflection (0.649), indicating that the verifier has learned to detect progressive physical dynamics such as phase transitions, light scattering, and color blending from intermediate DiT features. Categories with lower scores, such as Direct Radiation (0.320) and Flame Reaction (0.430), typically involve precise spatial reasoning or domain-specific material knowledge (e.g., shadow geometry, characteristic flame colors), which are underrepresented in the VideoPhy training distribution and suggest a clear avenue for improvement through targeted data augmentation. Notably, this category-dependent structure itself is an informative signal: it demonstrates that the verifier captures physically meaningful distinctions rather than relying on superficial quality cues. On Wan 2.1-14B, category-level variation is substantially reduced, with all means falling within $[0.42, 0.52]$, consistent with the distribution mismatch discussed above, and motivating backbone-matched training data as the primary path toward stronger cross-backbone transfer.



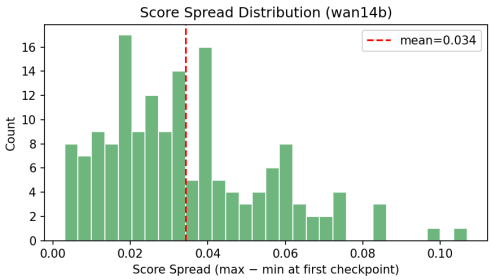
(a) CogVideoX-5B: kept vs. dropped



(b) Wan 2.1-14B: kept vs. dropped



(c) CogVideoX-5B: score spread



(d) Wan 2.1-14B: score spread

Figure S1: Verifier score analysis. Top: kept vs. dropped trajectory scores at checkpoints. Bottom: per-prompt score spread (max − min) at the first checkpoint. 5B shows meaningful separation; Wan shows near-complete overlap.

C More Qualitative Results

Figure S3 presents additional Wan 2.1-14B comparisons across mechanics, optics, thermal, and material prompts, each with a brief physical analysis of where the baseline violates the underlying law and how the guided result differs.

D Failure Cases

Figure S4 illustrates representative failure cases. These concentrate in material properties and chemical reaction prompts, which require domain-specific knowledge, such as characteristic flame colors, acid appearances, and fracture mechanics, that goes beyond the common-sense physical reasoning captured by our training data. These phenomena also involve dramatic transformations in material form (carbonization, shattering, phase changes), which current video diffusion models struggle to synthesize faithfully. When no candidate trajectory is physically correct, selection reaches a fundamental ceiling.

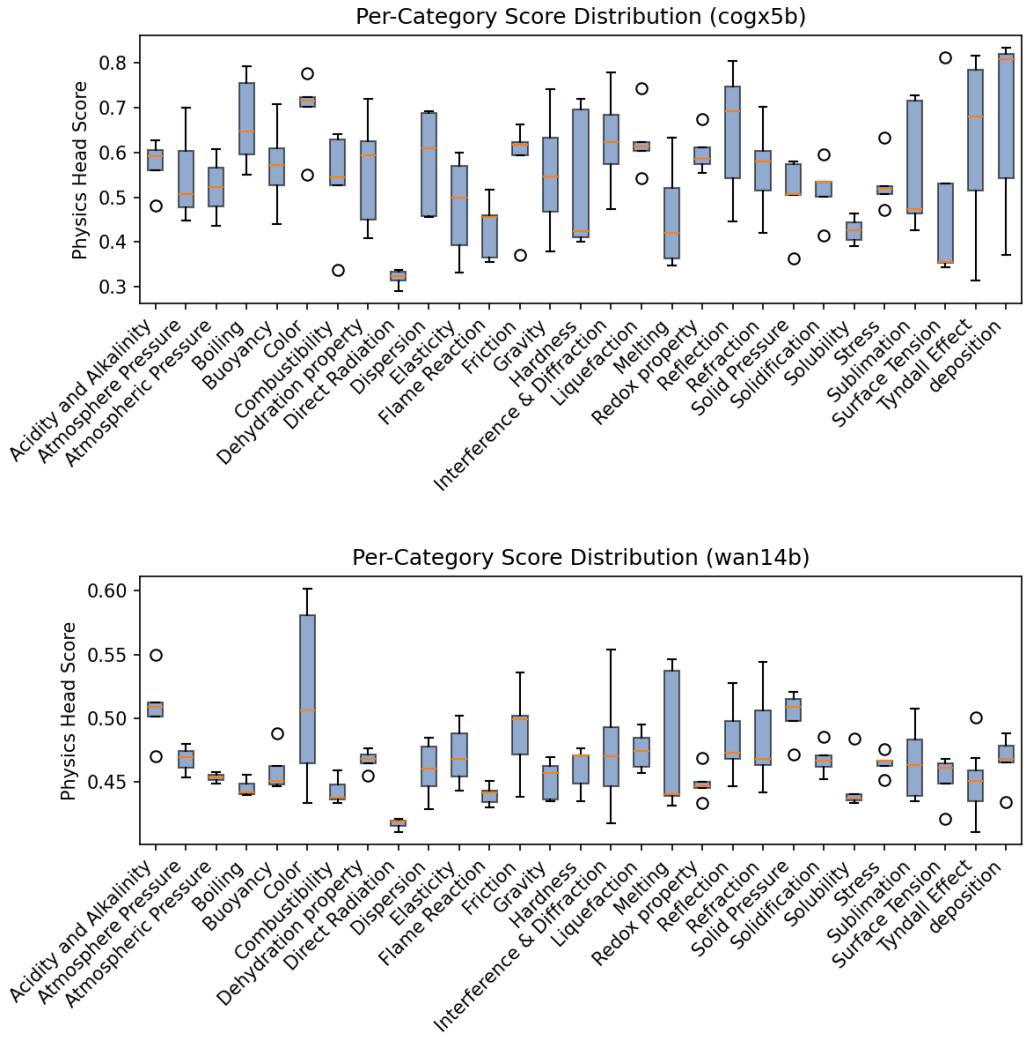


Figure S2: Per-category score distributions for CogVideoX-5B (top) and Wan 2.1-14B (bottom). On 5B, temporal phenomena (Color, Deposition, Boiling) score highest while spatial/material-specific phenomena (Direct Radiation, Flame Reaction) score lowest. On Wan, categories cluster in $[0.42, 0.52]$.

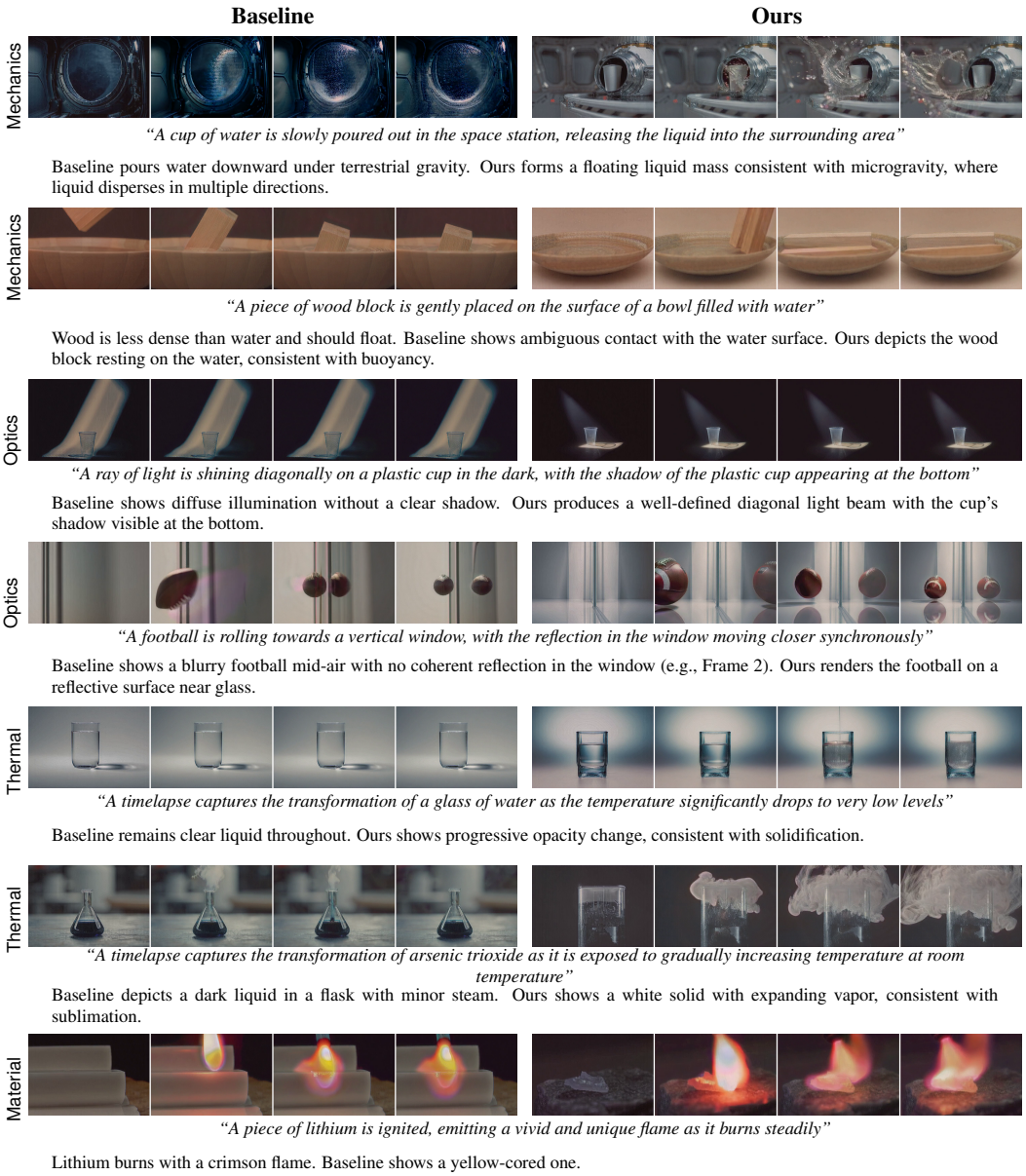
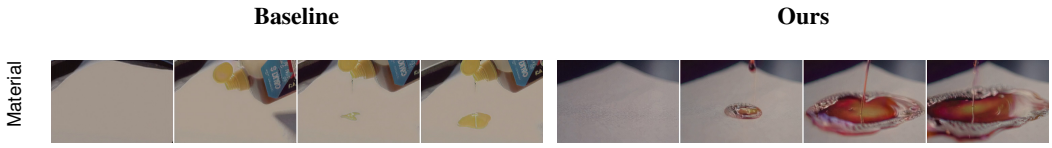


Figure S3: **Qualitative comparison** on PhyGenBench prompts (Wan-14B, 4 uniformly sampled frames per video). Each row shows baseline (left) vs. guided (right), the generation prompt, and a brief physical analysis.



"A timelapse captures the reaction as concentrated sulfuric acid is poured onto a sheet of paper"

Sulfuric acid is colorless and causes black carbonization. Baseline shows a yellow liquid with no reaction; ours shows a red liquid with paper degradation. Neither depicts the correct acid color or carbonization.



"A flimsy, brittle glass cup is propelled with force towards a solid, metallic surface, where it collides upon impact"

The cup should shatter into fragments. Both produce nearly identical results: the cup cracks but never exhibits realistic brittle fracture.

Figure S4: **Failure cases** (Wan-14B). Both rows show cases where all candidate trajectories are physically incorrect, leaving the verifier unable to select a better outcome.