RISK DECOMPOSITION FOR HALLUCINATION-AWARE VIDEO SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite achieving high visual fidelity, video models still suffer from physical hallucinations, limiting their reliability as world simulators. Existing methods leverage model-internal or external signals to improve physical plausibility, but their roles and limitations require a unified explanation. We propose a two-dimensional physical-hallucination risk framework that defines low-density sampling risk and high-density drifting risk. The first risk captures generations assigned low likelihood by the video model, for example, a ball suddenly teleporting despite the model favoring continuous motion, while the second risk captures physical inconsistencies in high-likelihood generations, for example, a ball bouncing progressively higher without additional energy. Our analysis explains how model-internal signals steer sampling toward higher-density regions of the learned distribution to address low-density sampling risk. However, higher model likelihood does not guarantee physical plausibility, motivating external guidance to address high-density drifting risk. Based on this framework, we develop Hallucination-Aware Sampler (HA-Sampler), which applies external physical guidance within latent refinement to generate physically plausible videos, and HA-Scorer, which derives a likelihood-based signal from the generator itself for evaluating generated videos using continuous normalizing flow (CNF), with coarse ODE integration and finite-difference divergence estimates. Experiments on Physics-IQ and VideoPhy benchmarks support our theoretical analysis and demonstrate that our methods outperform the compared sampling and scoring SOTA baselines.

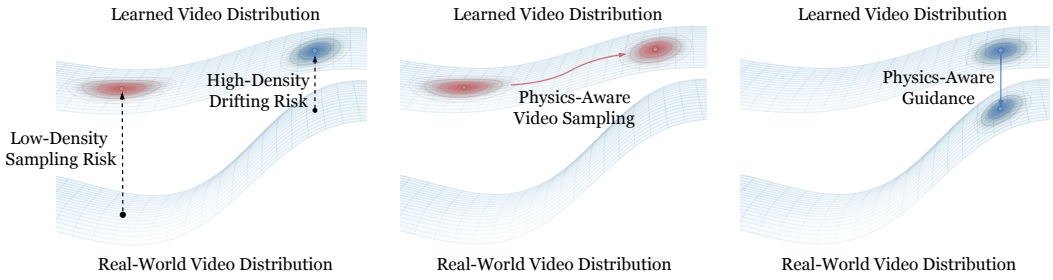


Figure 1: **Overview of our two-dimensional physics hallucination risk framework. Left:** Low-density sampling risk captures generations assigned low likelihood by the video model, while high-density drifting risk captures physical errors in high-likelihood generations. **Middle:** Model-internal physics-aware signals steer sampling toward higher-density regions of the learned distribution to reduce low-density sampling risk. **Right:** External physics-aware guidance helps correct the drifting between model-preferred generation and physical plausibility to address high-density drifting risk.

1 INTRODUCTION

Recent advances in visual fidelity and motion generation have highlighted the potential of video models as world simulators for downstream tasks, such as robotics (Yang et al., 2026). Such applications require accurate physical plausibility, yet physical hallucinations—such as disappearing objects and impossible interactions—undermine the reliability of generated videos (Bansal et al., 2025; Motamed et al., 2026). To address this, existing methods for video sampling can be catego-

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.

rized into two directions: one (Hassan et al., 2026; Yuan et al., 2026a) suggests that video models' own preferences contain useful information about physical plausibility and leverage model-internal signals (Jang et al., 2026; Cho et al., 2026) for refining, while the other uses external signals (Yuan et al., 2026b; Internò et al., 2026) to guide generation and filter physically unreal samples.

However, these raise a fundamental question: how do model-internal and external signals address physical hallucinations? We propose a unified framework that defines two-dimensional risks of physical hallucination to study and understand this question, as shown in Figure 1 (left). Low-density sampling risk captures the tendency to generate samples with low likelihood under the video model. For example, a model that favors continuous ball motion may nevertheless generate a low-likelihood trajectory in which the ball suddenly jumps from one location to another. On the contrary, high-density drifting risk captures physical inconsistencies in samples assigned high likelihood by the video model. For example, the model may favor a video of a ball bouncing progressively higher on a floor, despite no additional energy. The former reflects a deviation from the model's own preferences, whereas the latter exposes a drifting between those preferences and physical plausibility.

In our framework, we analyze how model-internal and external signals address the two risk dimensions, respectively. Model-internal signals steer video sampling toward the model's own preference region, which is higher-density regions of the learned video distribution, targeting the low-density sampling risk, as shown in Figure 1 (middle). However, these regions may drift from physical plausibility. External signal provides a physical guidance, forcing generations toward physically plausible videos from the model's own preference, addressing high-density drifting risk, as shown in Figure 1 (right). Our theory reveals the improvement by either signal alone is not sufficient.

Furthermore, we propose a hallucination-aware video sampler (HA-Sampler), and scorer (HA-Scorer). HA-Sampler generates physically plausible samples by integrating external physical guidance into model-based refinement, targeting both dimensions of hallucination risk. HA-Scorer derives a likelihood-based signal from the generator itself for evaluating generated videos, using continuous normalizing flow (CNF), with coarse ordinary differential equation (ODE) integration and finite-difference divergence estimates. Our extensive experiments show that HA-Sampler and HA-Scorer surpass state-of-the-art (SOTA) baselines, improving the physical plausibility of generated videos and helping select more physically plausible candidates, respectively. Our contributions can be summarized as follows:

1. We are the first to introduce a unified physical-hallucination risk framework that defines low-density sampling risk and high-density drifting risk.
2. We identify and demonstrate that video model-internal signals contain useful information related to physical plausibility that helps mitigate the first type of risk, but these signals are not always effective. External guidance can complement internal signals when they fall short by solving the second type of risk.
3. Based on our framework, we propose a hallucination-aware sampler, HA-Sampler, and scorer, HA-Scorer, for generating and scoring physically plausible videos, respectively. Our experiments show that these two proposed methods achieve the SOTA performance.

2 Related Work

2.1 Video generative models

Recent video generative models have achieved substantial progress in visual fidelity, temporal coherence, and controllability (Team Wan et al., 2025; Blattmann et al., 2023; Chen et al., 2025; HaCohen et al., 2025). They are commonly studied under two settings: text-to-video (T2V) (Chen et al., 2023; Singer et al., 2022; Hong et al., 2022; Tian et al., 2024), and image-to-video (I2V) (Blattmann et al., 2023; Team Wan et al., 2025; Hu et al., 2022; Yang et al., 2025; HaCohen et al., 2025). However, both T2V and I2V models still suffer from hallucination (Motamed et al., 2026; Li et al., 2025; Lin et al., 2025; Kang et al., 2025): generated videos are hard to preserve physical plausibility.

2.2 Physics-Aware Video Generation

Model-internal guidance. Methods based on model-internal guidance (Bai et al., 2025a; Wu et al., 2024; Jang et al., 2026; Shaulov et al., 2025) aim to enable models to refine themselves without external signal or verification. For video diffusion models, FreeInit (Wu et al., 2024) refines the

initial noise through an iterative denoise and re-noise procedure, which improves temporal consistency without training. Self-Refining Video Sampling (Jang et al., 2026) interprets flow-matching video generators as denoising autoencoders and proposes Predict-and-Perturb (P&P) steering the sampling trajectory toward more stable and self-consistent modes of the pretrained model. Tangential Amplifying Guidance (TAG) (Cho et al., 2026) amplifies the tangential component of each sampling update while preserving its radial component, steering samples toward regions of higher model density. Zigzag Diffusion Sampling (Bai et al., 2025a) proposes a sampling strategy that alternates denoising and inversion steps, allowing pretrained diffusion models to accumulate semantic information. FlowMo (Shaulov et al., 2025) improves motion coherence by extracting temporal representations from the model's own predictions and reducing patch-wise temporal variance during sampling. However, reinforcing the model's own preferences does not guarantee physical plausibility, as physically incorrect outcomes may still suffer from high model likelihood.

External guidance. Recent efforts on using external guidance introduce physical knowledge through additional supervision, pretrained models, or reward functions. PhysAlign (Xiong et al., 2026) uses simulation-derived annotations and video-feature alignment, while GeoPhys (Internò et al., 2026) assesses physical plausibility through the geometry of temporal feature trajectories from frozen image encoders. PhysRVG (Zhang et al., 2026) and PhyGDPO (Cai et al., 2025) improve physical consistency through physics-aware reinforcement learning and VLM-based preference optimization, respectively. VideoREPA (Zhang et al., 2025) transfers physical knowledge from video foundation models through relational alignment. At inference time, WMReward (Yuan et al., 2026b) uses a latent world model to score candidates and guide denoising. However, improving an external physical score does not necessarily increase model likelihood or reduce low-density sampling risk.

Hybrid strategies. Hybrid strategies coordinate model-internal signal with external guidance. Trust Sampling (Huang et al., 2024) monitors the model's predicted noise magnitude to terminate external optimization when samples depart from its estimated state manifold. TFG (Ye et al., 2024) further unifies guidance and recurrence within a common algorithmic design space. However, beyond coordinating these mechanisms, our framework explicitly distinguishes low-density sampling risk from high-density drifting risk, explaining the complementary roles and limitations of model-internal and external signals in mitigating physical hallucinations.

3 METHODOLOGY

3.1 NOTATIONS

Let $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}) \in \mathcal{Z}$ be the latent of a clean video $\mathbf{x} \in \mathcal{X}$, with frozen VAE encoder $\mathcal{E}$ and decoder $\mathcal{G}$. Conditioned on $\mathbf{c}$, $q^\star$ and q denote real-video and training latent distributions, while p_θ and $p_{\theta,\mathcal{S}}$ denote model distributions under ideal sampling and a practical sampler $\mathcal{S}$. We use the data-to-noise flow-matching path $\mathbf{Z}_t = (1-t)\mathbf{Z}_0 + t\mathbf{Z}_1$, $t \in [0,1]$, with $\mathbf{Z}_0 \sim q(\cdot \mid \mathbf{c})$, $\mathbf{Z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and target velocity $\mathbf{V} = \mathbf{Z}_1 - \mathbf{Z}_0$. Generation integrates the learned velocity field $v_\theta(\mathbf{z}_t, t, \mathbf{c})$ from $t = 1$ to $t = 0$, with clean endpoint prediction $\hat{\mathbf{z}}_{0|t} = \mathbf{z}_t - t\, v_\theta(\mathbf{z}_t, t, \mathbf{c})$.

3.2 PRELIMINARIES

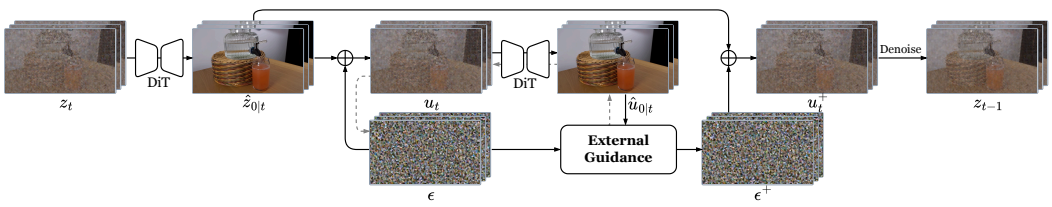
Mode interpolation. In prior work (Aithal et al., 2024), the ϵ-hallucination set is defined below.

$$H_\epsilon(q^*) = \{\mathbf{z}_0 : q^*(\mathbf{z}_0 \mid \mathbf{c}) \leq \epsilon\}. \tag{1}$$

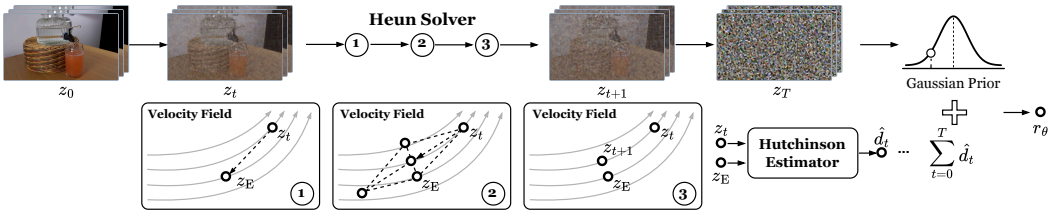
The parameter ϵ is taken to be vanishingly small. That is, if $\mathbf{z}_0 \in H_\epsilon(q)$, it lies outside of the support of the real data distribution, which means that the generated video was a hallucination case. The complement of the hallucination set is the ϵ-support set denoted by $S_\epsilon(q^*)$. Mode interpolation is a common phenomenon in diffusion models. This phenomenon occurs when the model generates samples that are interpolations of multiple modes in the ϵ-support set, such that the generated sample lies in the ϵ-hallucination set. For example, if $x, y \in S_\epsilon(q^*)$, then the model generates a sample $z = \alpha x + (1-\alpha)y$ for some $\alpha \in (0,1)$, and $z \in H_\epsilon(q^*)$.

3.3 TWO-DIMENSIONAL RISK

Directly quantifying hallucination in video generation is challenging, as it would require access to the unknown real-world conditional distribution q^*. We therefore aim to formulate a unified theoretical framework for characterizing hallucination risk in video models. Inspired by the prior



(a) HA-Sampler.



(b) HA-Scorer.

Figure 2: **Overview of our methods.** (a) HA-Sampler uses an external loss on the predicted clean video to guide the perturbation noise. The guided noise is combined with the original endpoint prediction before denoising continues. Grey dashed lines denote the flow of gradients. (b) HA-Scorer integrates the clean latent $\mathbf{z}_0$ to the noise endpoint $\mathbf{z}_1$ under the fixed base velocity field using a coarse Heun solver. Finite-difference Hutchinson estimator provides the divergence integral. In step 3, z_t and z_{t+1} fall on the same velocity line.

definition of the ϵ-hallucination set, we define the δ-low-density region as

$$D_{\theta,\delta}(\mathbf{c}) = \{\mathbf{z_0} : p_\theta(\mathbf{z}_0 \mid \mathbf{c}) \leq \delta\}. \tag{2}$$

Here, we decompose the ϵ-hallucination set into two parts:

$$H_\epsilon(q^* \mid \mathbf{c}) = \Big(H_\epsilon(q^* \mid \mathbf{c}) \cap D_{\theta,\delta}(\mathbf{c})\Big) \dot\cup \Big(H_\epsilon(q^* \mid \mathbf{c}) \cap \overline{D_{\theta,\delta}(\mathbf{c})}\Big) \coloneqq H_{\mathrm{I}} \dot\cup H_{\mathrm{II}}. \tag{3}$$

This decomposition distinguishes two fundamentally different hallucination regimes according to whether the generated sample falls also into the low-density region in the model learned distribution. H_{I} contains samples that have low-density in both q^* and p_θ. In contrast, H_{II} contains samples that are hallucinate under q^* but remain well supported by p_θ. Since q^* is inaccessible in practice, neither subset can be directly identified. However, low-density under p_θ is a necessary condition for membership in H_{I} and can be assessed using the model itself. This motivates us to characterize two layers of hallucination risk arising from sampling in low-density model regions and from drifting away from the real-world distribution despite high model density.

3.3.1 Low-Density Sampling Risk

The first layer of hallucination risk is motivated by the decomposed set H_{I}, which contains hallucinated samples that also fall into $D_{\theta,\delta}(\mathbf{c})$. Not every low-density sample is necessarily a hallucination. Nevertheless, mode interpolation provides an important mechanism through which generated samples may enter interpolated regions between learned modes (Aithal et al., 2024). Therefore, We treat low model density as an model internal indicator of hallucination risk.

Definition 1. ***(Low–Density Sampling Risk.)*** *Let $p_\theta(\cdot \mid \mathbf{c})$ denote the learned model distribution and $p_{\theta,\mathcal{S}}(\cdot \mid \mathbf{c})$ the output distribution induced by a practical sampler $\mathcal{S}$. We define the sample-wise low-density sampling risk as*

$$r_\theta(\mathbf{z} \mid \mathbf{c}) = -\log p_\theta(\mathbf{z} \mid \mathbf{c}). \tag{4}$$

Then, the low-density sampling risk of Sampler $\mathcal{S}$ is defined as

$$\mathcal{R}_\theta(\mathcal{S} \mid \mathbf{c}) = \mathbb{E}_{\mathbf{z}\sim p_{\theta,\mathcal{S}}(\cdot|\mathbf{c})}[r_\theta(\mathbf{z} \mid \mathbf{c})] = \mathbb{E}_{\mathbf{z}\sim p_{\theta,\mathcal{S}}(\cdot|\mathbf{c})}[-\log p_\theta(\mathbf{z} \mid \mathbf{c})]. \tag{5}$$

In our definition, the risk of a generated sample increases as the sample drops to the low-density region under p_θ. The sampler-wise risk measures the overall tendency to generate outputs that are weakly supported by the learned model.

3.3.2 HIGH-DENSITY DRIFTING RISK

Although reducing low-density sampling risk favors samples with higher density under p_θ, it only indicates that the model is more familiar with these samples. It does not guarantee that these samples are consistent with real-world dynamics. The learned distribution p_θ may still differ from the real-world physical distribution q^*, creating a gap that we refer to as distribution drifting. We therefore introduce a second layer of risk to characterize this drift.

Definition 2. ***(High–Density Drifting Risk.)*** *Let $\ell_\omega(\mathbf{x}, \mathbf{c}) \geq 0$ denote an external drifting loss, where a larger value indicates greater inconsistency between video $\mathbf{x}$ and real-world dynamics. For a sample with density above δ under p_θ, $\mathbf{z}_0 \notin D_{\theta,\delta}(\mathbf{c})$, we define its sample-wise high-density drifting risk as*

$$r_\omega(\mathbf{z}_0 \mid \mathbf{c}) = \ell_\omega\big(\mathcal{G}(\mathbf{z}_0), \mathbf{c}\big), \qquad \mathbf{z}_0 \notin D_{\theta,\delta}(\mathbf{c}). \tag{6}$$

The high-density drifting risk of sampler $\mathcal{S}$ is defined as the expected external loss conditioned on the generated sample lying in a region where its density under p_θ exceeds δ:

$$\mathcal{R}_\omega(\mathcal{S} \mid \mathbf{c}) = \mathbb{E}_{\mathbf{z}_0 \sim p_{\theta,\mathcal{S}}(\cdot|\mathbf{c})}\left[r_\omega(\mathbf{z}_0 \mid \mathbf{c}) \,|\, \mathbf{z}_0 \notin D_{\theta,\delta}(\mathbf{c})\right]. \tag{7}$$

This risk measures the average external loss among samples with relatively high density under p_θ. Conditioning on $\mathbf{z}_0 \notin D_{\theta,\delta}(\mathbf{c})$ follows the scope of H_{II} and measures the average severity of this distribution drifting among samples with relatively high model density. It does not imply that samples in $D_{\theta,\delta}(\mathbf{c})$ do not have external loss. Since q^* is inaccessible, the accuracy of this assessment depends on how well ℓ_ω captures this drift from real-world dynamics.

3.4 FROM TWO-DIMENSIONAL RISK TO PRACTICE

We use a first-order Euler solver as a reference sampler, denoted by $\mathcal{S}_{\text{E}}$. All sampler comparisons are conducted under the same condition $\mathbf{c}$. Previous work has shown that latent refinement exhibits mode-seeking behavior (Jang et al., 2026), while gradient guidance uses an external reward to steer generation toward more physically plausible outcomes (Yuan et al., 2026b). Mathematically, we interpret latent refinement through its idealized low-temperature Langevin counterpart at a fixed noise level, while gradient guidance approximates a reward-reweighted target distribution. We provide a detailed mathematical analysis in Appendix A. Figure 3 shows that latent refinement used in P&P reduces low-density sampling risk in more cases than high-density drifting risk, while WMReward shows the reverse pattern. Under our framework, latent refinement addresses low-density sampling risk by favoring regions of higher model density, while gradient guidance provides a way to address high-density drifting through the external guidance. These complementary roles motivate the incorporation of gradient guidance into the refinement process to address both risks within each iteration.

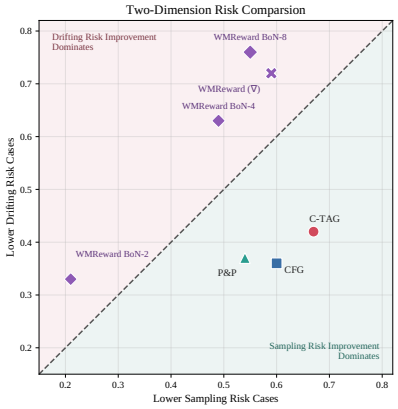
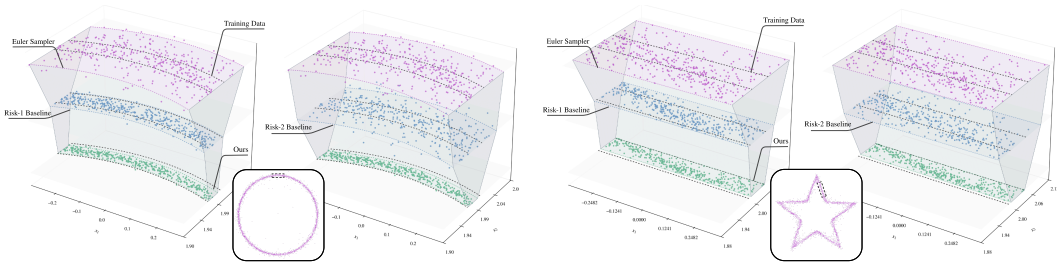


Figure 3: **Performance comparison across the two-dimensional risk.** Each axis reports the fraction of 100 cases with lower corresponding risk than the Euler sampler.

3.4.1 HALLUCINATION-AWARE VIDEO SAMPLING

Building on our two-layer hallucination risk framework, we introduce HA-Sampler, a training-free sampler designed to address low-density sampling risk and high-density drifting risk within the same refinement process. At each iteration, gradient guidance adjusts the perturbation noise, and the generative model re-predicts the clean endpoint from the guided perturbation. The sampling procedure is illustrated in Figure 2(a). We further illustrate this design through two-dimensional toy experiments with circle- and star-shaped data distributions in Figure 4. We compare HA-Sampler with Euler sampling and baselines that target either low-density sampling risk or high-density drifting risk alone. In the displayed regions, our samples are more tightly concentrated around the training-data reference curves than those from Euler or either baseline.

Sampling procedure. At noise level t, we perform K_t refinement iterations, starting from the current predicted clean endpoint $\hat{\mathbf{z}}_{0|t}$. Each iteration samples perturbation noise $\boldsymbol{\epsilon}$ to construct $\mathbf{u}_t$



(a) Circle distribution. (b) Star distribution.

Figure 4: **Sampling results on 2D toy distributions.** The stacked layers compare Euler sampler, P&P (risk-1 baseline), guidance based on the negative log-likelihood (NLL) under the true data distribution (risk-2 baseline), and HA-Sampler. The Euler sampler uses a large number of sampling steps to reduce discretization error. Purple reference curves indicate the training-data geometry.

Method	Physics-IQ Metrics				Score ↑
	Spatial IoU ↑	Spatiotemporal IoU ↑	Weighted Spatial IoU ↑	MSE Score ↑	
SANA-Video2.0 (Chen et al., 2026)	0.2209	0.1986	0.1309	0.0630	15.33
CogVideoX-5B (Yang et al., 2025)	0.3686	0.3806	0.2286	0.2971	31.87
Wan2.2-TI2V-5B (Team Wan et al., 2025)	0.3295	0.2442	0.1727	0.3074	26.34
SkyReels-V2-I2V (Chen et al., 2025)	0.2717	0.2642	0.1917	0.1084	20.90
SkyReels-V2-I2V	0.2717	0.2642	0.1917	0.1084	20.90
+ P&P (Jang et al., 2026)	0.2974	0.2685	0.2157	**0.1365**	22.95
+ WMReward (∇) (Yuan et al., 2026b)	0.2742	0.2640	0.1939	0.1100	21.05
+ **HA-Sampler** (ours)	**0.3132**	**0.2750**	**0.2342**	0.1289	**23.78**
Wan2.2-TI2V-5B	0.3295	**0.2442**	0.1727	**0.3074**	26.34
+ P&P (Jang et al., 2026)	0.3796	0.2322	0.2109	0.2916	27.86
+ WMReward (∇) (Yuan et al., 2026b)	0.3310	0.2427	0.1732	0.3072	26.35
+ **HA-Sampler** (ours)	**0.3927**	0.2360	**0.2243**	0.2846	**28.44**

Table 1: **Comparison of different samplers on Physics-IQ.** We report the Physics-IQ Verified metrics and overall score. The lower blocks compare inference-time methods using Wan2.2-TI2V-5B and SkyReels-V2-I2V. ∇ denotes gradient guidance of WMReward.

and predict its clean endpoint $\hat{\mathbf{u}}_{0|t}$. We compute the external loss $\mathcal{L} = \ell_\omega(\mathcal{G}(\hat{\mathbf{u}}_{0|t}), \mathbf{c})$ and its gradient with respect to $\boldsymbol{\epsilon}$. After root mean square (RMS) normalization, we update the noise with step size η_t. We then construct $\mathbf{u}_t^+$ with the updated noise and re-evaluate the model to obtain $\hat{\mathbf{z}}_{0|t}^+$, which becomes the input to the next iteration. Algorithm 1 in Appendix B presents one refinement iteration.

3.4.2 Hallucination-Aware Scoring

To estimate hallucination risk r_θ in Definition 1, we integrate the fixed base ODE $\frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t} = v_\theta(\mathbf{z_t}, t, \mathbf{c})$ from the clean latent $\mathbf{z}_0$ to its noise endpoint $\mathbf{z}_1$. The continuous normalizing flow (CNF) change-of-variables formula (Chen et al., 2018) gives the following

$$-\log p_\theta(\mathbf{z}_0 \mid \mathbf{c}) = -\log \varphi(\mathbf{z}_1) - \int_0^1 \nabla_{\mathbf{z}_t} \cdot v_\theta(\mathbf{z}_t, t, \mathbf{c}) dt, \tag{8}$$

where φ is the standard Gaussian density. The first term is the negative log-likelihood (NLL) under the Gaussian prior, whereas the integral accounts for volume changes along the flow. Computing the divergence is expensive for high-dimensional video latent. We estimate the divergence using Hutchinson's estimator (Hutchinson, 1989), with forward finite differences to avoid backpropagation through the video model:

$$\hat{d}_t = \frac{1}{P}\sum_{p=1}^{P} \mathbf{r}_p^\top \mathbf{J}_{\theta,t}\mathbf{r}_p \approx \frac{1}{P\delta_t}\sum_{p=1}^{P} \mathbf{r}_p^\top \big[v_\theta(\mathbf{z}_t + \delta_t \mathbf{r}_p, t, \mathbf{c}) - v_\theta(\mathbf{z}_t, t, \mathbf{c})\big]. \tag{9}$$

Here, $\mathbf{J}_{\theta,t}$ is the Jacobian of v_θ with respect to $\mathbf{z}_t$. The random probes $\mathbf{r}_p$ are sampled independently and reused across integration steps. The perturbation scale δ_t is computed from δ_{rel} and the latent root mean square. We use Heun's second-order method (Karras et al., 2022) on a coarse time grid

Method	Physics-IQ Metrics				Score ↑
	Spatial IoU ↑	Spatiotemporal IoU ↑	Weighted Spatial IoU ↑	MSE Score ↑	
		Base sampler			
Random	0.3372	0.1828	0.2418	0.2936	26.38
Qwen3-VL (Bai et al., 2025b)	0.3449	**0.2572**	0.2161	0.3091	28.18
Gemma-4 (Gemma Team, 2026)	0.3380	0.2392	0.1847	0.3002	26.55
WMReward (Yuan et al., 2026b)	0.3175	0.1972	**0.3647**	**0.3410**	30.51
GeoPhys (Internò et al., 2026)	0.3278	0.1782	0.2307	0.2900	25.67
HA-Scorer (ours)	**0.3515**	0.2172	0.3312	0.3353	**30.88**
		P&P sampler			
Random	0.3647	0.2308	0.2001	0.2852	27.02
Qwen3-VL (Bai et al., 2025b)	0.3910	0.2368	0.2197	0.3020	28.74
Gemma-4 (Gemma Team, 2026)	0.3455	0.2257	0.1897	0.2888	26.24
WMReward (Yuan et al., 2026b)	0.3681	**0.3025**	0.2304	0.3246	30.64
GeoPhys (Internò et al., 2026)	0.3757	0.2049	0.2058	0.2925	26.97
HA-Scorer (ours)	**0.3917**	0.2384	**0.2910**	**0.3349**	**31.40**
		HA-Sampler			
Random	0.3681	0.2020	0.2325	0.2846	27.18
Qwen3-VL (Bai et al., 2025b)	0.3777	0.2385	0.2178	0.3057	28.49
Gemma-4 (Gemma Team, 2026)	0.3652	0.2354	0.2095	0.2958	27.65
WMReward (Yuan et al., 2026b)	0.3761	0.2303	**0.3294**	0.3315	31.68
GeoPhys (Internò et al., 2026)	0.3686	0.2011	0.2190	0.2907	26.98
HA-Scorer (ours)	**0.3965**	**0.2443**	0.3127	**0.3364**	**32.25**

Table 2: **Results of scorer methods on Physics-IQ with Best-of-N generation.** Random denotes selecting a candidate uniformly from the $N = 8$ candidates pool. Candidate generation uses Wan2.2-TI2V-5B.

to compute the inversion trajectory and accumulate the finite-difference divergence estimates. We refer to this evaluation as HA-Scorer. An overview of HA-Scorer is shown in Figure 2 (b), with the complete procedure given in Algorithm 2 (Appendix B).

4 Experiments

4.1 Experimental Setup

Implementation details. We conducted all experiments on flow-matching-based video generative models, including Wan2.2-TI2V-5B and SkyReels-V2-I2V-1.3B. For VLM-based scoring, we use Qwen3-VL-8B-Instruct (Bai et al., 2025b) and Gemma-4-E4B-it (Gemma Team, 2026). Details of the scoring procedure and results with other VLMs are provided in Appendix D.1. In practice, we use the WMReward surprise score (Yuan et al., 2026b) as the external loss. For HA-Scorer, the default settings are $H = 8$ Heun steps, $P = 2$ Hutchinson probes, and a relative perturbation scale of $\delta_{\text{rel}} = 0.0125$. We compare scoring methods using a best-of-N (BoN) protocol, where each method selects its top-ranked candidate from $N = 8$ video candidates generated per input condition. Comparisons with $N \in \{2, 4, 8\}$ are provided in Appendix D.2 to examine how performance changes with the number of candidates. All experiments are conducted on the RTXA6000 platform.

Baselines and benchmarks. For sampling, we compare with the base model, P&P (Jang et al., 2026), and WMReward (∇) (Yuan et al., 2026b). We also include other physics-aware methods in the T2V evaluation. For candidate selection, we compare against WMReward (Yuan et al., 2026b), GeoPhys (Internò et al., 2026). We choose these baselines for their competitive and SOTA performance. We evaluate I2V generation on Physics-IQ (Rädsch et al., 2026) using the Verified metrics and T2V generation on VideoPhy (Bansal et al., 2025). Details are provided in Appendix C.

4.2 Comparison with Existing Methods

4.2.1 Quantitative Results

Sampler comparison. Table 1 compares inference-time sampling methods on Physics-IQ Benchmark. On Wan2.2-TI2V-5B, HA-Samplerachieves the highest overall score among compared sam-

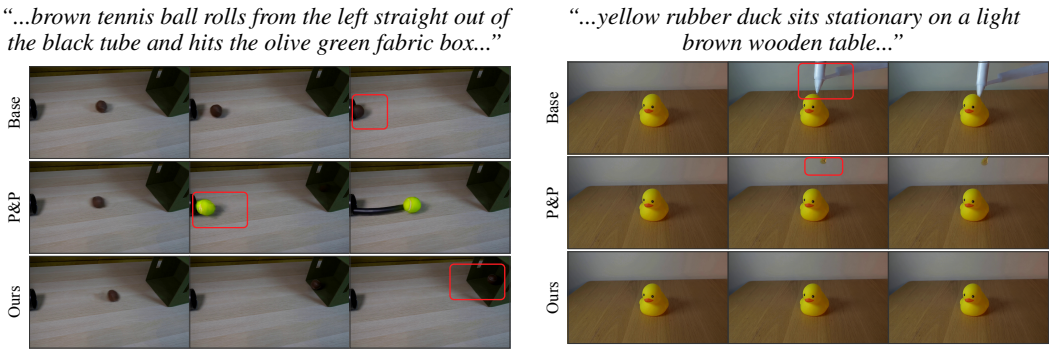


Figure 5: **Qualitative comparison of samplers on Physics-IQ.** We compare sampling methods in a dynamic collision scene (left) and a static scene (right). HA-Sampler produces more physically plausible results than the base sampler and P&P in both cases.

Method	Solid–Solid ↑	Solid–Fluid ↑	Fluid–Fluid ↑	Overall ↑
General video models				
Wan2.1-T2V-1.3B (Team Wan et al., 2025)	0.18	0.24	0.27	0.22
SanaVideo2.0 (Chen et al., 2026)	**0.27**	0.23	0.33	0.26
Wan2.2-TI2V-5B (Team Wan et al., 2025)	0.21	**0.32**	0.25	0.26
Physics-aware methods				
DreamWorld (Tan et al., 2026)	0.16	0.29	0.31	0.24
VPT (Zheng et al., 2026)	0.17	0.24	0.33	0.22
P&P (Jang et al., 2026)	0.22	0.23	0.33	0.24
WMReward (∇) (Yuan et al., 2026b)	0.21	**0.32**	0.24	0.26
HA-Sampler (ours)	**0.27**	0.30	**0.35**	**0.30**

Table 3: **Comparison of different samplers on VideoPhy.** We report category-wise and overall Physical Commonsense (PC). Higher scores are better. Our method uses Wan2.2-TI2V-5B as the base model. ∇ denotes gradient guidance of WMReward.

plers, exceeding the baseline by **+2.10**, and leads in Spatial IoU and Weighted Spatial IoU. On SkyReels-V2-I2V, our method also achieves the highest overall score, exceeding the baseline by **+2.88**. These results suggest that latent refinement and gradient guidance complement each other, producing consistent improvements over either method alone across both backbones. Table 3 reports results on VideoPhy. HA-Sampler achieves the highest overall PC score among the compared physics-aware methods. We further compare the two hallucination risks across samplers in Appendix E.

Scorer comparison. Table 2 compares scorer methods for best-of-N selection with $N = 8$ using Wan2.2-TI2V-5B. HA-Scorer achieves the highest overall score under all three samplers. Compared with random selection, HA-Scorer improves the overall score by **+4.50**, **+4.38**, and **+5.07** under the base, P&P, and HA-Sampler, respectively. Our scorer also leads in Spatial IoU across all three samplers and ranks first in three of the four metrics under both P&P and HA-Sampler. These gains show that our scorer effectively selects more physically plausible videos and remains beneficial even when candidate generation is improved through HA-Sampler.

4.2.2 Qualitative Results

Sampler comparison. Figure 5 compares sampling methods in dynamic and static scenes. In the left case, the base sampler moves the ball back toward the tube, while P&P changes the ball's color from brown to yellow and introduces an additional tube. HA-Sampler preserves the appearance of the original ball and generates a trajectory into the box. In the right case, both baselines introduce extraneous objects above the duck, whereas HA-Sampler preserves the stationary scene.

Scorer comparison. Figure 6 compares videos selected by different scorers from the same base-model candidate pool for each prompt. In the left case, additional objects appear in the videos selected by Qwen3-VL and WMReward, and the ball also disappears in the video selected by WMReward. By contrast, HA-Scorer selects a video in which the ball remains visible and rolls toward the duck. In the right case, both Qwen3-VL and WMReward select videos in which the ball flies

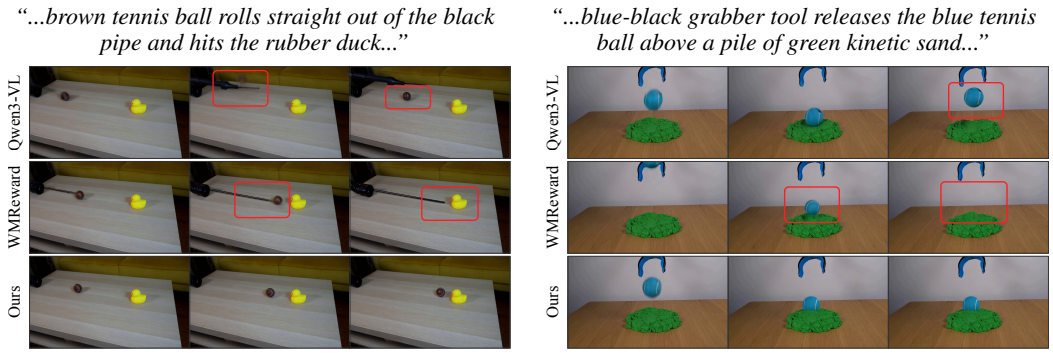


Figure 6: **Qualitative comparison of candidates selection on Physics-IQ.** For each prompt, all scorer methods select from the same candidate pool generated by the base model. Each row shows three frames capturing key stages of motion in the video.

Setting	Physics-IQ Metrics				Physics-IQ Score ↑	VideoPhy PC ↑
	Spatial IoU ↑	Spatiotemporal IoU ↑	Weighted Spatial IoU ↑	MSE Score ↑		
w/o both	0.3295	**0.2442**	0.1727	**0.3074**	26.34	<u>0.26</u>
w/ Latent Refinement	<u>0.3796</u>	0.2322	<u>0.2109</u>	<u>0.2916</u>	<u>27.86</u>	0.24
w/ Physics Guidance	0.3252	<u>0.2435</u>	0.1680	**0.3074**	26.10	<u>0.26</u>
HA-Sampler (ours)	**0.3927**	0.2360	**0.2243**	0.2846	**28.44**	**0.30**

Table 4: **Component ablation of HA-Sampler.** We evaluate latent refinement and physics guidance individually and together; w/o both uses neither component. All settings use the same backbone, prompts, initial frames, and paired seeds, without candidate selection. The best and second-best results in each column are bold and underlined, respectively.

upward unnaturally. By contrast, HA-Scorer selects a video in which the ball lands on the sand and remains at rest. More qualitative results are provided in Appendix G.

4.3 Ablation Studies

Table 4 evaluates the contributions of latent refinement and physics guidance. Removing either component lowers the overall Physics-IQ score and VideoPhy PC. Guidance alone performs similarly to the base sampler, whereas adding it to latent refinement raises the overall score. These results support the complementary roles of refinement and guidance, demonstrating the effectiveness of our sampling mechanism. Furthermore, our sensitivity analysis shows that HA-Scorer maintains stable performance across the tested hyperparameter settings; see Appendix F for details.

5 Discussion

Our approach may improve physical plausibility at the expense of generation diversity. HA-Sampler and HA-Scorer favor high-density regions of the learned distribution, potentially suppressing rare but physically valid outputs. This trade-off may limit applications that require diverse video generation. Preserving the diversity of valid outputs while reducing physical hallucinations remains an important direction for future work. Furthermore, we discuss potential applications in synthetic-data generation for robotics and training guidance in Appendix H.

6 Conclusion

We proposed a unified framework that distinguishes low-density sampling risk from high-density drifting risk in video generation. Our analysis explains the complementary roles and limitations of model-internal and external signals: internal signals favor higher-likelihood generations, while external physical guidance helps address physical errors that the model itself assigns high likelihood. Based on this framework, we proposed HA-Sampler for video sampling and HA-Scorer for likelihood-based video evaluation. Experiments demonstrate improvements in generating and selecting physically plausible videos. Beyond inference-time sampling, these risk measures could help select synthetic training videos and assess video models during post-training.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking and writing refinement. We have not used generative AI tools for any other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have not used generative AI tools for any additional tasks with recommended disclosure. We have reviewed all AI-assisted work. All AI-assisted text was carefully reviewed and revised by the authors to ensure accuracy and consistency with the intended meaning. 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 studies training-free methods for estimating and reducing physical hallucination risk in video generation. Our experiments use publicly released pretrained models and public benchmarks, including Physics-IQ and VideoPhy. Video generation models can be misused to produce misleading or harmful content; our methods operate at inference time without retraining the base models and remain subject to their content-related risks and limitations. We encourage responsible use of our methods in accordance with the licenses and usage policies of the underlying models and datasets.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. Detailed pseudocode for our algorithms is provided in Appendix B. The experimental setup in Section 4.1 specifies the hyperparameters, models, and benchmarks used in our experiments. The VLM-based scoring procedure and prompts are described in Appendix D.1. The code will be publicly released upon acceptance.

REFERENCES

Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in neural information processing systems*, 37:134614–134644, 2024.

Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag Diffusion Sampling: Diffusion models can self-improve via self-reflection. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=MKvQH1ekeY.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b. doi: 10.48550/arXiv.2511.21631.

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. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fce2d8a485746f76aac7b5650db2679d-Abstract-Conference.html.

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.

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. *arXiv preprint arXiv:2512.24551*, 2025. doi: 10.48550/arXiv.2512.24551.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Junsong Chen, Jincheng Yu, Yitong Li, Shuchen Xue, Haozhe Liu, Jingyu Xin, Yuyang Zhao, Tian Ye, Zhangjie Wu, Zian Wang, Daquan Zhou, Ping Luo, Song Han, and Enze Xie. SANA-Video 2.0: Hybrid linear attention with attention residuals for efficient video generation, 2026. URL https://arxiv.org/abs/2607.21553.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Hyunmin Cho, Donghoon Ahn, Susung Hong, Jee Eun Kim, Seungryong Kim, and Kyong Hwan Jin. TAG: Tangential amplifying guidance for hallucination-resistant sampling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=1XJqPy2LhA.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025.

Mariam Hassan, Kaouther Messaoud, Wuyang Li, and Alexandre Alahi. Proprio: Latent self-scoring and inference-time refinement for physically plausible video generation. *arXiv preprint arXiv:2605.28230*, 2026. URL https://arxiv.org/abs/2605.28230.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18219–18228, 2022.

William Huang, Yifeng Jiang, Tom Van Wouwe, and C Karen Liu. Constrained diffusion with trust sampling. *Advances in Neural Information Processing Systems*, 37:93849–93873, 2024.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806.

Christian Internò, Alexander Pondaven, Habon Issa, Fabio Pizzati, Francesco Pinto, Markus Olhofer, Ivan Laptev, Philip Torr, Eero P. Simoncelli, Barbara Hammer, and David Klindt. GeoPhys: The geometry of physical plausibility, 2026. URL https://arxiv.org/abs/2606.20707.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025. URL https://arxiv.org/abs/2505.12705.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-refining video sampling. *arXiv preprint arXiv:2601.18577*, 2026. doi: 10.48550/arXiv.2601.18577.

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, 2025. URL https://arxiv.org/abs/2411.02385.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, et al. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 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 *Computer Vision – ECCV 2024*, pp. 366–384. Springer, 2024. doi: 10.1007/978-3-031-72673-6_20.

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, March 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Tim Rädsch, Yuki M. Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-IQ Verified, 2026. URL https://arxiv.org/abs/2606.18943.

Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. FlowMo: Variance-based flow guidance for coherent motion in video generation. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2506.01144. URL https://openreview.net/forum?id=PDgagkX9zj.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Boming Tan, Xiangdong Zhang, Ning Liao, Yuqing Zhang, Shaofeng Zhang, Xue Yang, Qi Fan, and Yanyong Zhang. DreamWorld: Unified world modeling in video generation, 2026. URL https://arxiv.org/abs/2603.00466.

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. doi: 10.48550/arXiv.2503.20314.

Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. Videotetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems*, 37:29489–29513, 2024.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging initialization gap in video diffusion models. In *Computer Vision – ECCV 2024*, pp. 378–394. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72646-0_22.

Zhexiao Xiong, Yizhi Song, Liu He, Wei Xiong, Yu Yuan, Feng Qiao, and Nathan Jacobs. PhysAlign: Physics-coherent image-to-video generation through feature and 3d representation alignment. *arXiv preprint arXiv:2603.13770*, 2026. doi: 10.48550/arXiv.2603.13770.

Jiazhi Yang, Kunyang Lin, Jinwei Li, Wencong Zhang, Tianwei Lin, Longyan Wu, Zhizhong Su, Hao Zhao, Ya-Qin Zhang, Li Chen, Ping Luo, Xiangyu Yue, and Hongyang Li. RISE: Self-improving robot policy with compositional world model. *arXiv preprint arXiv:2602.11075*, 2026. doi: 10.48550/arXiv.2602.11075. URL https://arxiv.org/abs/2602.11075.

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*, 2025.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37:22370–22417, 2024.

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*, 2026a. URL https://openreview.net/forum?id=6UJf6B8RZ8.

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. doi: 10.48550/arXiv.2601.10553.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. PhysRVG: Physics-aware unified reinforcement learning for video generative models. *arXiv preprint arXiv:2601.11087*, 2026. doi: 10.48550/arXiv.2601.11087.

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*, 2025.

Guangting Zheng, Haojing Chen, Hao Li, Jingtao Zhang, Zhen Yang, Xiaosong Jia, Xue Yang, Shaofeng Zhang, and Yanyong Zhang. Enhancing video physical consistency via role-aware joint training and modality-decoupled denoising, 2026. URL https://arxiv.org/abs/2607.04653.

A Mathematical Analysis

This section analyzes latent refinement and gradient guidance within our two-layer risk framework. We interpret latent refinement through an idealized low-temperature Langevin process and gradient guidance through distribution reweighting based on the external loss. These analyzes motivate combining the two mechanisms in HA-Sampler to address low-density sampling risk and high-density drifting risk.

(i) Latent Refinement. At a fixed noise level t, latent refinement predicts a clean endpoint and adds fresh Gaussian noise to return it to the same noise level. It repeats this prediction–perturbation cycle before advancing to the next denoising step (Jang et al., 2026). Under a Bayes-optimal endpoint predictor for the Gaussian noising process, Tweedie's formula (Efron, 2011) gives

$$\mathbf{z}_t^{k+1} = \mathbf{z}_t^k + \sigma_t^2 \nabla_{\mathbf{z}_t} \log p_{\theta,t}(\mathbf{z}_t^k \mid \mathbf{c}) + \sigma_t \boldsymbol{\epsilon}_{k+1}, \qquad \boldsymbol{\epsilon}_{k+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

where k indexes refinement iterations. The transition combines a step along the model score with a Gaussian perturbation. Its continuous Langevin counterpart has temperature $T = 1/2$ and stationary density

$$\pi_t(\mathbf{z}_t \mid \mathbf{c}) = \frac{p_{\theta,t}(\mathbf{z}_t \mid \mathbf{c})^2}{Z_t(\mathbf{c})}, \tag{11}$$

where $Z_t(\mathbf{c})$ is the normalization constant. Squaring the density increases the relative weight of higher-density regions, and this distribution satisfies

$$\mathbb{E}_{\mathbf{z}_t \sim \pi_t(\cdot|\mathbf{c})}[-\log p_{\theta,t}(\mathbf{z}_t \mid \mathbf{c})] \leq \mathbb{E}_{\mathbf{z}_t \sim p_{\theta,t}(\cdot|\mathbf{c})}[-\log p_{\theta,t}(\mathbf{z}_t \mid \mathbf{c})]. \tag{12}$$

The expected negative log-density here is the noisy-marginal counterpart of the risk in Definition 1 which explains how latent refinement can reduce low-density sampling risk at a fixed noise level.

(ii) Gradient Guidance. Gradient guidance predicts a clean endpoint during denoising, decodes it into a video, and evaluates the drifting risk r_ω through the external loss. The resulting gradient adjusts the sampling update toward outcomes that the evaluator considers more physically plausible (Yuan et al., 2026b). The ideal reweighted target associated with this risk is

$$p_{\theta,\lambda}(\mathbf{z}_0 \mid \mathbf{c}) = \frac{p_\theta(\mathbf{z}_0 \mid \mathbf{c}) \exp[-\lambda r_\omega(\mathbf{z}_0 \mid \mathbf{c})]}{Z_\lambda(\mathbf{c})}, \tag{13}$$

where $\lambda \geq 0$ controls guidance strength and $Z_\lambda(\mathbf{c})$ is the normalizing constant. The exponential factor increases the relative weight of samples with lower drifting risk. Let $p_{\theta,\lambda,t}$ denote its marginal under the same noising process. The corresponding score is approximated as

$$\nabla_{\mathbf{z}_t} \log p_{\theta,\lambda,t}(\mathbf{z}_t \mid \mathbf{c}) \approx \nabla_{\mathbf{z}_t} \log p_{\theta,t}(\mathbf{z}_t \mid \mathbf{c}) - \lambda \nabla_{\mathbf{z}_t} r_\omega(\hat{\mathbf{z}}_{0|t} \mid \mathbf{c}). \tag{14}$$

Prior work has reported limited gains in physical plausibility from gradient guidance alone in video generation (Yuan et al., 2026b). Our analysis suggests that guidance can be more effective when low-density sampling risk is reduced or kept under control.

B Algorithms

Algorithm 1 presents one refinement iteration of HA-Sampler, where gradient guidance updates the perturbation noise before the model re-predicts the clean endpoint. For Wan2.2-TI2V-5B, we use 20 denoising steps, with three refinement iterations at steps 1–2 and one at steps 3–6. For SkyReels-V2-I2V-1.3B, we use 50 denoising steps, with three refinement iterations at steps 3–7 and one at steps 8–17. Algorithm 2 details HA-Scorer, which estimates the negative log-likelihood per latent dimension using Heun integration and finite-difference divergence estimates.

Algorithm 1 HA-Sampler

Input: Predicted clean endpoint $\hat{\mathbf{z}}_{0|t}$, noise level t, guidance step size η_t, velocity field v_θ, VAE decoder $\mathcal{G}$, external loss ℓ_ω, condition $\mathbf{c}$.

Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{u}_t \leftarrow (1-t)\hat{\mathbf{z}}_{0|t} - t\boldsymbol{\epsilon}$
$\mathbf{v} \leftarrow v_\theta(\mathbf{u}_t, t, \mathbf{c})$
$\hat{\mathbf{u}}_{0|t} \leftarrow \mathbf{u}_t - t\mathbf{v}$ ▷ Probe endpoint
$\mathcal{L} \leftarrow \ell_\omega\big(\mathcal{G}(\hat{\mathbf{u}}_{0|t}), \mathbf{c}\big)$
$\mathbf{g} \leftarrow \nabla_{\boldsymbol{\epsilon}}\mathcal{L}$ ▷ Noise gradient
$\mathbf{g} \leftarrow \mathrm{RMSNormalize}(\mathbf{g})$
$\boldsymbol{\epsilon}^+ \leftarrow \boldsymbol{\epsilon} - \eta_t\mathbf{g}$ ▷ Guided noise
$\mathbf{u}_t^+ \leftarrow (1-t)\hat{\mathbf{z}}_{0|t} - t\boldsymbol{\epsilon}^+$
$\mathbf{v}^+ \leftarrow v_\theta(\mathbf{u}_t^+, t, \mathbf{c})$ ▷ Re-predict
$\hat{\mathbf{z}}_{0|t}^+ \leftarrow \mathbf{u}_t^+ - t\mathbf{v}^+$ ▷ Refined endpoint
return $\hat{\mathbf{z}}_{0|t}^+, \mathbf{u}_t^+, \mathbf{v}^+$

Algorithm 2 HA-Scorer

Input: Clean endpoint $\mathbf{z}_0$, base velocity field v_θ, condition $\mathbf{c}$, time grid $0 = t_0 < \cdots < t_H = 1$, probe count P, relative perturbation scale δ_{rel}.

Sample independent probes $\mathbf{r}_1, \ldots, \mathbf{r}_P$ ▷ Fixed across steps
$\mathbf{z}_{t_0} \leftarrow \mathbf{z}_0$, $A \leftarrow 0$, $D \leftarrow \dim(\mathbf{z}_0)$
for $j = 0, \ldots, H-1$ **do**
  $h_j \leftarrow t_{j+1} - t_j$
  $(\mathbf{v}_j, \hat{d}_j) \leftarrow \textsc{Evaluate}(\mathbf{z}_{t_j}, t_j)$
  $\mathbf{z}_{\mathrm{E}} \leftarrow \mathbf{z}_{t_j} + h_j\mathbf{v}_j$ ▷ Euler predictor
  $(\mathbf{v}_j^{\mathrm{E}}, \hat{d}_j^{\mathrm{E}}) \leftarrow \textsc{Evaluate}(\mathbf{z}_{\mathrm{E}}, t_{j+1})$
  $A \leftarrow A + \frac{h_j}{2}(\hat{d}_j + \hat{d}_j^{\mathrm{E}})$ ▷ Divergence integral
  $\mathbf{z}_{t_{j+1}} \leftarrow \mathbf{z}_{t_j} + \frac{h_j}{2}(\mathbf{v}_j + \mathbf{v}_j^{\mathrm{E}})$ ▷ Heun update
$\hat{r}_\theta \leftarrow \frac{\|\mathbf{z}_{t_H}\|_2^2}{2} + \frac{D}{2}\log(2\pi) - A$ ▷ Estimated NLL
return $\hat{r}_\theta / D$ ▷ NLL per dimension

function $\textsc{Evaluate}(\mathbf{z}, t)$
  $\mathbf{v} \leftarrow v_\theta(\mathbf{z}, t, \mathbf{c})$
  $\delta \leftarrow \delta_{\mathrm{rel}} \max\big(\|\mathbf{z}\|_2/\sqrt{D},\, 0.1\big)$
  $\hat{d} \leftarrow \frac{1}{P\delta}\sum_{p=1}^{P} \mathbf{r}_p^\top\big[v_\theta(\mathbf{z} + \delta\mathbf{r}_p, t, \mathbf{c}) - \mathbf{v}\big]$ ▷ Forward FD
  return $\mathbf{v}, \hat{d}$

C Experimental Details

We describe the baselines, benchmarks, and evaluation metrics used in our experiments. Sampling methods and best-of-N video selection are evaluated separately.

C.1 Baselines

C.1.1 Sampler Baselines

Base model. The pretrained video generator uses the reference Euler sampler with the official classifier-free guidance (CFG) settings, without additional latent refinement or physics guidance.

P&P. (Jang et al., 2026) P&P interprets flow-matching video generators as denoising autoencoders and repeatedly predicts and perturbs latent states at a fixed noise level to favor stable model modes.

WMReward (∇). (Yuan et al., 2026b) We use its gradient-guidance variant, which guides sampling using prediction surprise from a pretrained latent world model as the external loss.

DreamWorld. (Tan et al., 2026) DreamWorld jointly models video latents with motion, spatial, and semantic features extracted by pretrained models. It uses constraint annealing during training and multi-source inner-guidance at inference.

VPT. (Zheng et al., 2026) VPT fine-tunes video generators with optical flow and entity-role maps, using modality-decoupled denoising and auxiliary loss annealing. It further applies cross-step auto-guidance at inference to improve physical consistency.

C.1.2 Scorer Baselines

WMReward. (Yuan et al., 2026b) We use prediction surprise from its pretrained latent world model to rank candidate videos, selecting the candidate with the lowest surprise.

GeoPhys. (Internò et al., 2026) GeoPhys evaluates physical plausibility from the geometry of temporal feature trajectories extracted by frozen image encoders. We reimplement its scoring method with the recommended DINOv3 backbone.

VLM scorers. We evaluate Qwen3-VL-8B-Instruct (Bai et al., 2025b), Gemma-4-E4B-it (Gemma Team, 2026), and Qwen3.5-9B (Qwen Team, 2026). All three use the same binary answer-likelihood scoring procedure described in Appendix D.1.

For video selection, all scorers rank the same candidate pool for each condition, with random selection serving as the reference.

C.2 Metrics

Spatial IoU. Compares temporally aggregated motion masks to measure whether motion occurs in the correct spatial regions.

Spatiotemporal IoU. Compares frame-wise motion masks to assess both spatial and temporal alignment.

Weighted Spatial IoU. Compares weighted motion maps to assess the location and magnitude of motion.

Mean Squared Error (MSE). Measures pixel-level differences between generated and reference continuations.

C.3 Benchmarks

Physics-IQ Verified. We evaluate image-to-video generation using the Physics-IQ Verified metrics (Rädsch et al., 2026). The benchmark compares generated continuations with recordings of real-world physical experiments. Its four underlying metrics are described below.

We use the Verified aggregation, which normalizes each metric by the physical variation of the corresponding sample, clips each normalized subscore to $[0, 1]$, and averages equally across metrics and samples. The MSE score is transformed so that higher values indicate better performance.

VideoPhy. We evaluate text-to-video generation on the 344-prompt VideoPhy test split (Bansal et al., 2025). We report semantic adherence (SA), which measures agreement with the text prompt, and physical commonsense (PC), which assesses whether the generated interactions follow physical laws.

D Details of Best-of-N Selection

D.1 VLM Scoring and Selection

Following (Jang et al., 2025), we provide the VLM with a generated video and its input description, and ask it to assess physical plausibility with a binary response: `1` for physically plausible and `0` otherwise. Following (Lin et al., 2024), we use the log-probability of the positive answer token `1` as

the score and select the highest-scoring candidate. The prompt used for VLM scoring is provided below.

```
[Does the video show good physics dynamics and showcase
a good alignment with the physical world for the prompt
"description"?  Please be a strict judge.  If it breaks the
laws of physics, please answer 0.  Answer 0 for No or 1 for
Yes.  Reply only 0 or 1.]
```

We use Qwen3-VL-8B-Instruct (Bai et al., 2025b) as the primary VLM. We also evaluate Gemma-4-E4B-it (Gemma Team, 2026) and Qwen3.5-9B (Qwen Team, 2026), using the same scoring procedure and candidate pools. Table 5 compares best-of-N selection results across the three VLMs to examine how the evaluator choice affects the relative performance of the scorer methods. Across all three VLMs, the physical evaluation scores of the selected videos do not consistently improve as N increases. This suggests that these VLMs still struggle to reliably identify physical errors and select more physically plausible videos.

Sampler	BoN	Qwen3-VL	Gemma-4	Qwen3.5
Base sampler	2	**27.73**	26.48	27.27
	4	**28.64**	26.61	28.12
	8	28.18	26.55	**29.42**
P&P sampler	2	28.33	27.08	**28.69**
	4	28.42	26.08	**28.71**
	8	**28.74**	26.24	28.22
HA-Sampler	2	28.59	28.57	**29.09**
	4	28.16	26.66	**29.21**
	8	28.49	27.65	**29.17**

Table 5: **Comparison of VLM-based selection across candidate pool sizes.** We report Physics-IQ Verified scores for candidates selected by different VLMs at $N \in \{2, 4, 8\}$. Higher scores indicate better performance. The highest score in each row is shown in bold.

D.2 Detailed Results

Table 6 compare scorer methods with $N \in \{2, 4, 8\}$ candidates generated by the base sampler, P&P, and HA-Sampler. HA-Scorer achieves the highest Physics-IQ Verified score in every setting, with scores increasing consistently across the tested candidate budgets for all three samplers. Combined with HA-Sampler, it achieves the best result in each budget.

Sampler	BoN	Random	WMReward	GeoPhys	Qwen3-VL	HA-Scorer (ours)
Base sampler	2	27.74	<u>28.84</u>	26.57	27.73	**28.99**
	4	27.45	<u>30.75</u>	26.50	28.64	**30.77**
	8	26.38	<u>30.51</u>	25.67	28.18	**30.88**
P&P sampler	2	27.28	<u>28.79</u>	27.08	28.33	**29.38**
	4	27.09	<u>29.88</u>	26.34	28.42	**30.48**
	8	27.02	<u>30.64</u>	26.97	28.74	**31.40**
HA-Sampler (ours)	2	27.54	<u>29.05</u>	28.49	28.59	**29.53**
	4	27.04	<u>30.00</u>	27.39	28.16	**30.69**
	8	27.18	<u>31.68</u>	26.98	28.49	**32.25**

Table 6: **Best-of-N results with different candidate pool sizes.** We report Physics-IQ Verified scores for selection methods using $N \in \{2, 4, 8\}$ candidates from each sampler. Random denotes uniform random selection. Higher scores indicate better performance. The best and second-best scores in each row are bold and underlined, respectively.

E HALLUCINATION RISK ACROSS SAMPLERS

We compare both risk measures across samplers on 100 cases in Table 7 and Table 8. Low-density sampling risk is estimated by HA-Scorer on the final clean latents, while mean WMReward surprise is used as a proxy for high-density drifting risk.

Low-density sampling risk. C-TAG, P&P, CFG, and HA-Sampler all reduce the estimated risk relative to Euler without CFG, whereas WMReward (∇) increases it. Unlike WMReward (∇), HA-Sampler reduces low-density sampling risk while applying gradient guidance.

Sampler	Risk Decomposition			ΔRisk/dim
	Prior term	Volume term	Risk/dim ↓	
Euler	1.10503	-2.54120	-1.43617	0.00000
C-TAG ($\lambda = 1.25$)	1.22623	-2.76701	-1.54079	-0.10462
P&P	1.21343	-2.68283	-1.46940	-0.03323
CFG	1.21803	-2.70666	-1.48863	-0.05246
WMReward (∇)	1.22506	-2.21974	-0.99468	+0.44149
HA-Sampler (ours)	1.21706	-2.72160	-1.50454	-0.06837

Table 7: **Low-density sampling risk on 100 cases.** The prior and volume terms sum to the low-density sampling risk. All quantities are reported per latent dimension. ΔRisk is measured relative to the paired Euler baseline without classifier-free guidance. We use $H = 8$ Heun steps and $P = 2$ shared Hutchinson probes. ∇ denotes gradient guidance.

High-density drifting risk. Relative to the base sampler with CFG $= 5$, P&P slightly increases mean WMReward surprise, while WMReward (∇) and HA-Sampler reduce it. HA-Sampler achieves the lowest mean risk among the evaluated methods. Together, these results show that HA-Sampler reduces both low-density sampling risk and high-density drifting risk.

Sampler	Risk ↓	ΔRisk
Base (CFG $= 5$)	0.414941	0.000000
P&P	0.415273	+0.000332
WMReward (∇)	0.414746	-0.000195
HA-Sampler (ours)	0.414180	-0.000761

Table 8: **High-density drifting risk on 100 cases.** We report mean WMReward surprise as a proxy for high-density drifting risk. ΔRisk is measured relative to the base sampler with CFG $= 5$. ∇ denotes gradient guidance.

F HA-SCORER SENSITIVITY

F.1 SENSITIVITY TO H AND p

Table 9 examines how the Heun step count H and Hutchinson probe count p affect HA-Scorer. More steps or probes reduce the mean absolute error (MAE) relative to the reference NLL estimates, while rank correlations with the reference estimates remain close to one across the tested settings. These results indicate that the rankings produced by HA-Scorer are relatively insensitive to H and p within the tested range. Table 10 further shows that varying H and P within the tested range has little effect on candidate selection or physical evaluation results.

H	NFE	MAE ↓	ρ ↑
100 cases			
4	24	0.067996	0.99600
8	48	0.003014	0.99957
16	96	Ref.	1.00000
25 case subset			
16	96	0.001210	1.00000
32	192	Ref.	1.00000

(a) **Heun integration steps.**

p	NFE	MAE ↓	ρ ↑
100 cases			
1	32	0.000548	0.99994
2	48	0.000322	0.99994
4	80	Ref.	1.00000
25 case subset			
4	80	0.000256	1.00000
8	144	Ref.	1.00000

(b) **Hutchinson probe count.**

Table 9: **Sensitivity of HA-Scorer to integration steps and probe count.** We evaluate the full 100-case set and a fixed stratified 25-case subset. We fix $p = 2$ in (a), $H = 8$ in (b), and $\delta_{\text{rel}} = 0.0125$ throughout. ρ denotes the Spearman rank correlation coefficient. NFE counts model forward evaluations per case.

(a) Probe count ($H = 8$)

Sampler	Physics-IQ Score ↑		ΔScore	Agreement
	$P = 1$	$P = 2$		
Base sampler	30.88	30.88	+0.00	198/198
P&P sampler	31.28	31.40	+0.12	197/198
HA-Sampler	32.25	32.25	+0.00	198/198

(b) Integration steps ($P = 1$)

Sampler	Physics-IQ Score ↑		ΔScore	Agreement
	$H = 4$	$H = 8$		
HA-Sampler	32.17	32.25	+0.08	191/198

Table 10: **Effect of estimation settings on HA-Scorer.** We compare Physics-IQ Verified scores using different numbers of Heun steps (H) and Hutchinson probes (P). ΔScore is the change from the left setting to the right. Agreement counts prompts with the same selected candidate.

F.2 Sensitivity to δ_{rel}

Table 11 examines the sensitivity of HA-Scorer to δ_{rel} by comparing forward finite-difference estimates with VJP-based estimates. The default setting produces rankings close to the VJP reference, while the smallest and largest tested perturbation scales lead to larger estimation errors.

Estimator / δ_{rel}	Mean Risk/dim	Bias vs. VJP	MAE ↓	Pearson ↑	Spearman ↑	Kendall ↑	Mean rank shift ↓
VJP	-1.48701	–	–	–	–	–	–
Forward finite differences							
0.0050	-1.51387	-0.02686	0.08914	0.95765	0.93638	0.80889	7.30
0.0100	-1.47179	+0.01523	0.02618	0.99493	0.99123	0.93374	2.52
0.0125	-1.46672	+0.02030	0.02890	0.99456	0.99191	0.93051	2.70
0.0250	-1.47240	+0.01461	0.01739	0.99846	0.99735	0.96485	1.40
0.0500	-1.40883	+0.07818	0.07818	0.99800	0.99740	0.96687	1.40
0.1000	-1.17969	+0.30732	0.30732	0.97516	0.97786	0.87475	4.54

Table 11: **Finite-difference accuracy across perturbation scales.** We compare forward finite differences with VJP-based estimates. δ_{rel} denotes the relative perturbation scale. Bias and MAE are computed relative to VJP, while correlations and mean rank shift measure agreement across evaluated samples.

G More Qualitative Results

G.1 Qualitative Results on Physics-IQ

Figure 8 and Figure 9 compare our method with the baselines on Physics-IQ, focusing on generated motion and object interactions. Compared with the baselines, our method produces more physically plausible motion and object interactions in these examples. These observations are consistent with the quantitative results and support combining HA-Sampler for generation with HA-Scorer for candidate selection.

G.2 Qualitative Results on VideoPhy

Figure 7 compares T2V results with P&P and WMReward (∇). In the left case, HA-Sampler produces a cleaner contact between brush and canvas. In the right case, it produces a more natural interaction with the water surface. These results illustrate HA-Sampler has improvements in physical details. Additional qualitative comparisons on VideoPhy are provided in Figure 10.

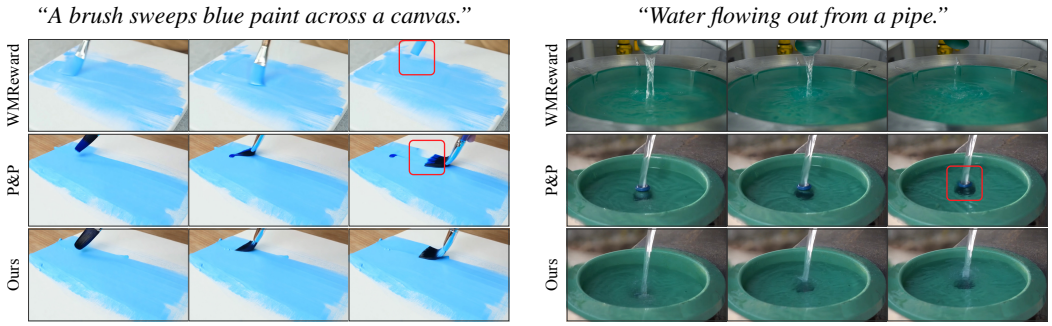


Figure 7: **Qualitative comparison of T2V generation on VideoPhy.** For each prompt, we compare different samplers using three representative frames from each generated video. HA-Sampler improves physical plausibility in both cases and helps correct inconsistencies in the generated interactions.

H Applications

Reliable data generator. HA-Sampler supports more reliable synthetic-data generation for critical downstream tasks, such as robotics. HA-Sampler improves the physical plausibility of generated motion and interactions, while HA-Scorer provides a physics-aware criterion for selecting candidate videos.

Training-free signal for training. HA-Scorer could also provide a diagnostic signal for monitoring video models during training without training an additional evaluator. A possible protocol is to generate videos from successive checkpoints under fixed conditions and sampling settings, then score them using our HA-Scorer. The resulting likelihood-based risk could help track hallucination-related failure patterns and guide checkpoint selection.

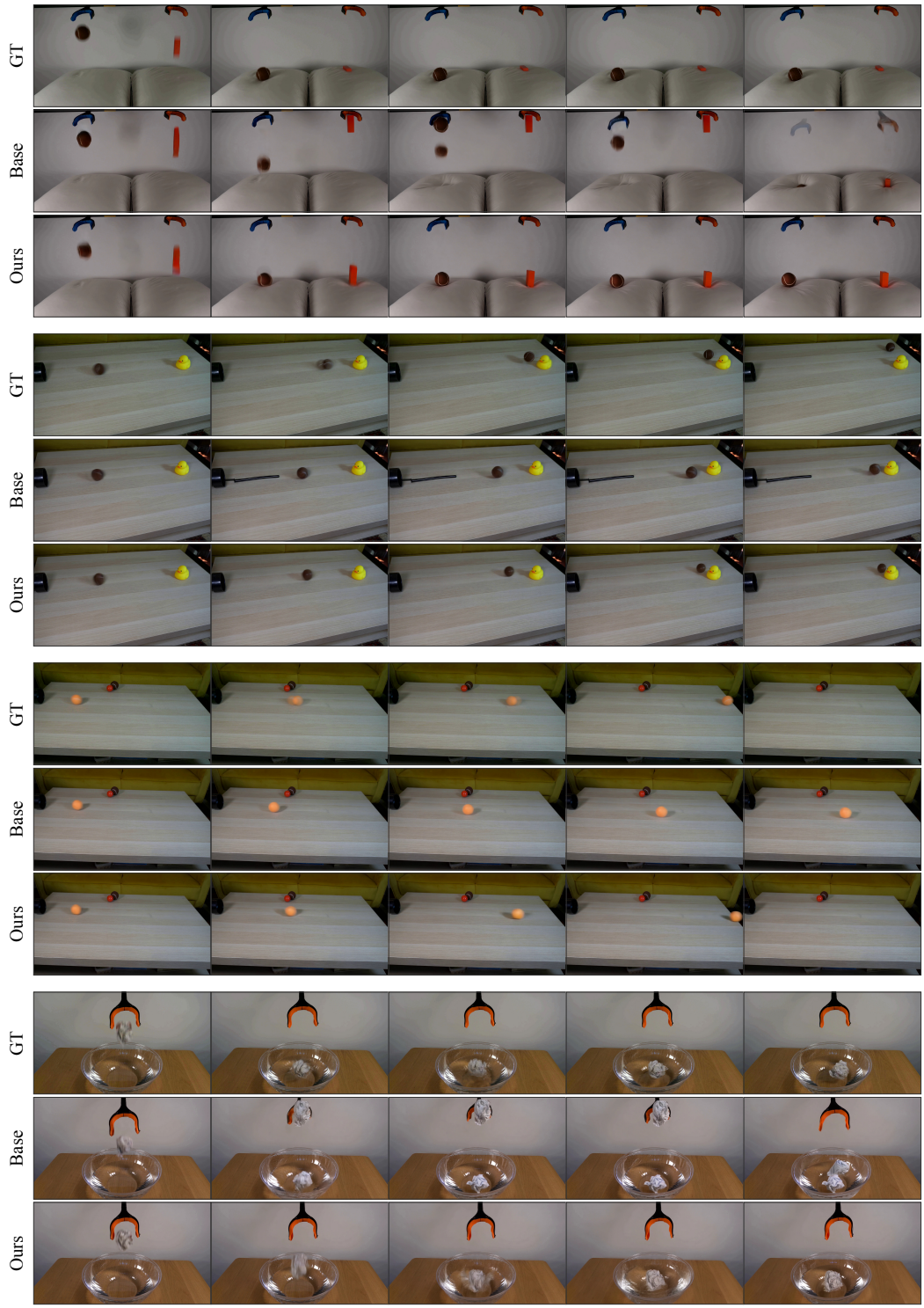


Figure 8: **Qualitative comparisons on Physics-IQ.** Each example shows GT, Base, and Ours from top to bottom, with five frames arranged chronologically. Ours denotes videos generated with HA-Sampler and selected using HA-Scorer.

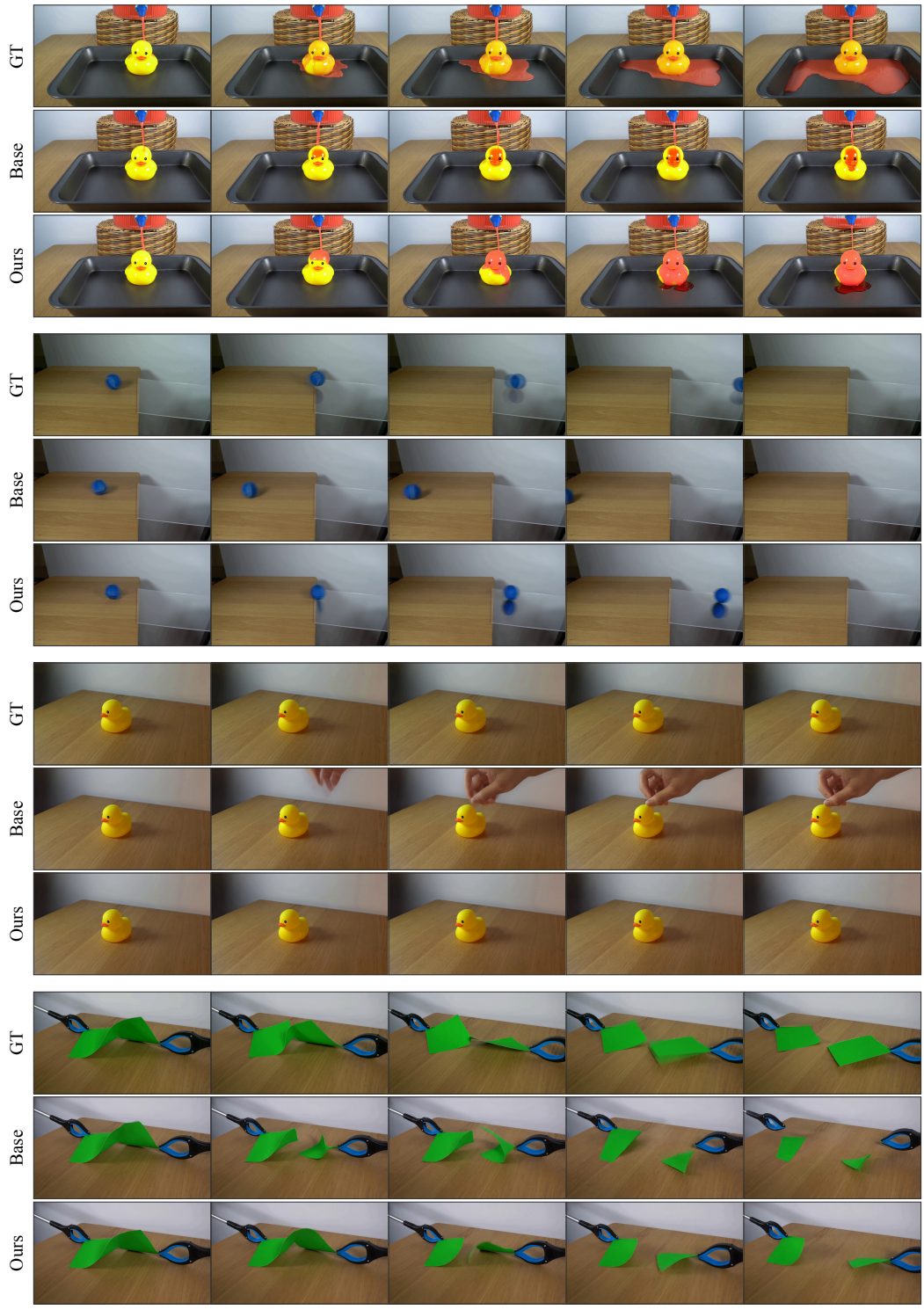


Figure 9: **Additional qualitative comparisons on Physics-IQ.** The layout follows Figure 8.

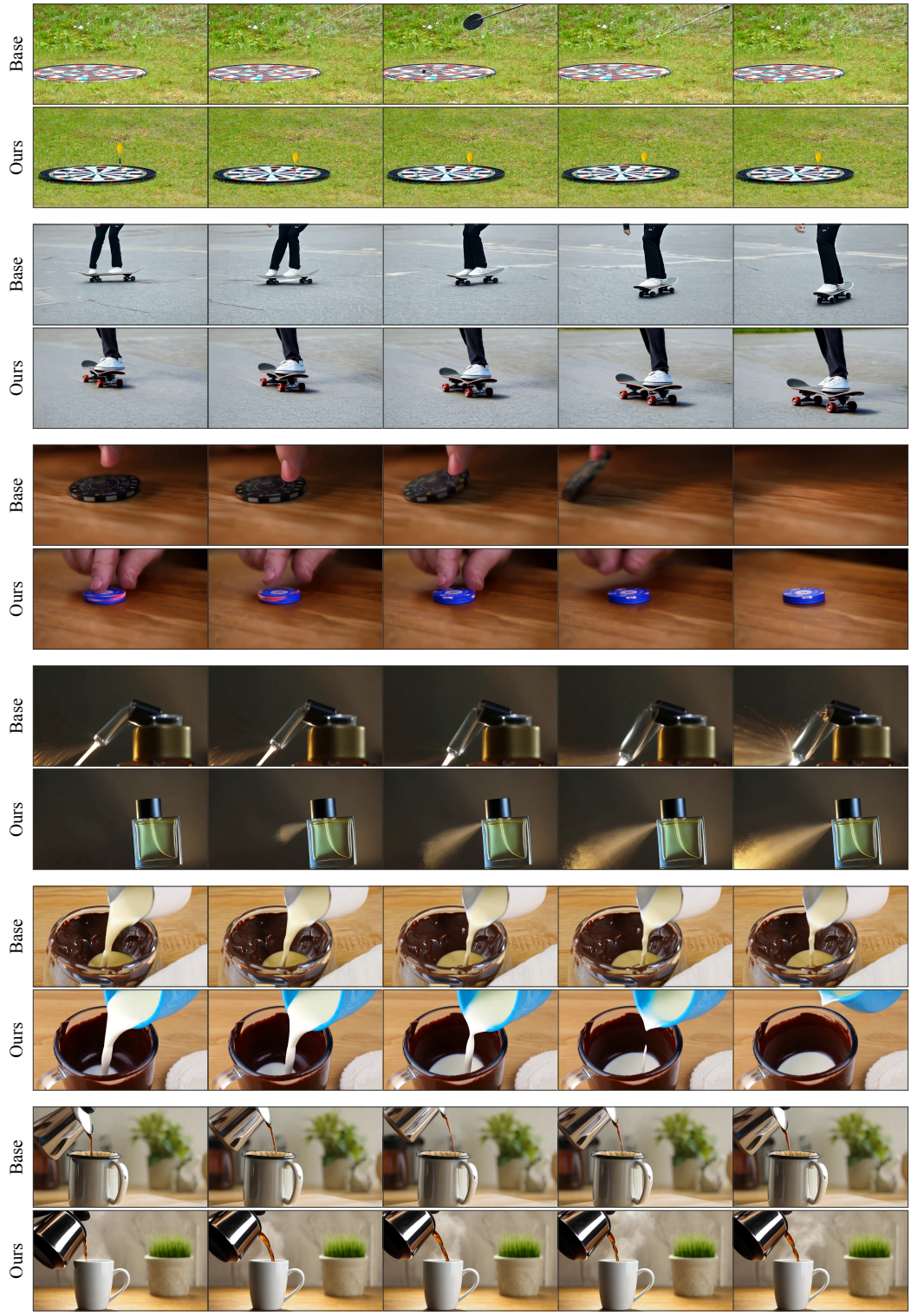


Figure 10: **Qualitative comparisons on VideoPhy.** For each prompt, we show five frames from videos produced by the base model and HA-Sampler (Ours).