This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Inference-time Physics Alignment of Video Generative Models with Latent World Models

Jianhao Yuan[1,2,*] Xiaofeng Zhang[1,3,4,‡] Felix Friedrich[1,‡]
Nicolas Beltran-Velez[1,5,*,‡] Melissa Hall[1] Reyhane Askari-Hemmat[1]
Xiaochuang Han[1] Nicolas Ballas[1] Michal Drozdzal[1,†] Adriana Romero-Soriano[1,3,6,7,†]

[1]FAIR, Meta Superintelligence Labs [2]University of Oxford [3]Mila – Québec AI Institute
[4]Université de Montréal [5]Columbia University [6]McGill University [7]Canada CIFAR AI Chair

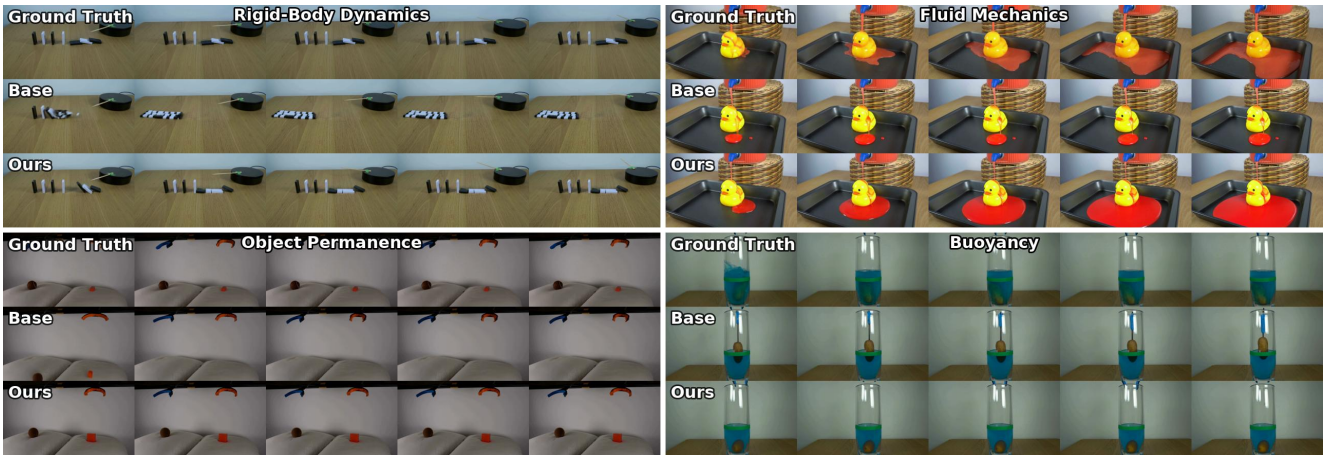


Figure 1. We improve the physics plausibility of video generation by aligning a pre-trained diffusion model with a latent world model at inference time. Using a reward derived from latent world models, we perform search on guided denoising trajectories to sample from a tilted physics plausible distribution. Compared to the baseline (middle row of each quadrant), our generation (bottom) adheres more closely to real-world physics (top), exhibiting smoother temporal continuity, more accurate solid interactions, and improved fluid behavior.

Abstract

State-of-the-art video generative models produce promising visual content yet often violate basic physics principles, limiting their utility. While some attribute this deficiency to insufficient physics understanding from pre-training, we find that the shortfall in physics plausibility also stems from suboptimal inference strategies. We therefore introduce `WMReward` *and treat improving physics plausibility of video generation as an inference-time alignment problem. In particular, we leverage the strong physics prior of a latent world model (here, VJEPA-2) as a reward to search and steer multiple candidate denoising trajectories, enabling scaling test-time compute for better generation performance. Empirically, our approach substantially improves physics plausibility across image-conditioned, multiframe-conditioned, and text-conditioned generation settings, with validation from human preference study. Notably, in the* ICCV 2025 Perception Test PhysicsIQ Challenge*, we achieve a final score of* 62.64%*, winning first place and outperforming the previous state of the art by* 7.42%*. Our work demonstrates the viability of using latent world models to improve physics plausibility of video generation, beyond this specific instantiation or parameterization. Code is available at* https://github.com/facebookresearch/WMReward.

1. Introduction

State-of-the-art video generative models [9, 14, 31, 45, 48, 52, 59] show remarkable capabilities in generating visually pleasing videos. Yet, the progress in generation quality has been hindered by the limited physics understanding of these

*Work done at Meta ‡Equal contribution †Joint last author
†Correspondence: {adrianars, mdrozdzal}@meta.com, jianhaoyuan@robots.ox.ac.uk

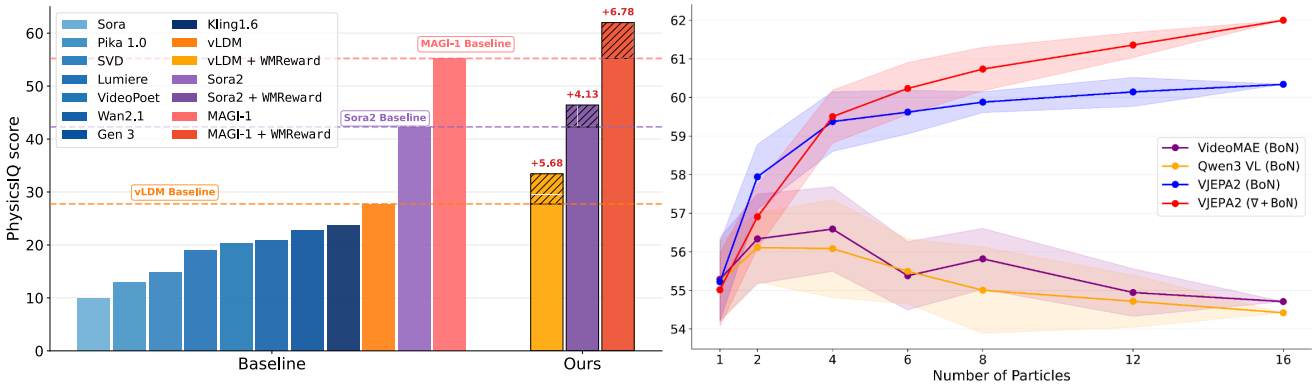


Figure 2. **Left:** We achieve the new state-of-the-art on PhysicsIQ benchmark, improving physics plausibility in both single (I2V) and multiframe (V2V) conditioned video generation by a considerable margin. **Right:** Our latent world model reward `WMReward` substantially outperforms VLM and other vision foundation model-based reward signals under BoN search; the proposed ∇+`BoN` sampling strategy further improves the scaling effect.

models [30, 44, 69], resulting in physically implausible videos [7, 8]. Ensuring physics correctness in video generation is not only crucial to increase user satisfaction, but also for reliable world modeling [34] and downstream applications like robotics [66] or autonomous driving [25].

Prior work attributes the observed physical implausibility to the pre-training stage of video generative models [30], which relies on minimizing pixel or feature-level reconstruction errors. Following this hypothesis, substantial work has focused on improving pre-training or post-training of video generative models by injecting physics information [16, 17, 38, 70, 71, 73]. By contrast, another line of work assumes that physically plausible videos may be found in the manifold learned by the generative model and therefore has focused on devising inference-time methods to improve physics. This is an underexplored line of research, with two contributions relying on vision-language models (VLMs) to rewrite prompts [64] and plan motion [67] when a motion-controllable video generative model is available [15]. However, in the image generation literature, alternatives to prompt rewriting [21, 72] have been extensively explored to search over the generated data manifold with the goal of boosting performance at inference-time. In particular, reward models have been used to perform candidate selection with search over random seeds [37, 43, 74] and to guide the model's sampling process towards high utility generations [2, 3, 20, 24, 27, 68]. Despite these advances, existing work has not explored the use of *off-the-shelf* reward models to improve the physics plausibility of video generation at inference time.

Recently, latent world models have demonstrated strong capabilities in physics understanding [13, 23, 42]. A *latent world model* is a predictive model that encodes high-dimensional observations (*e.g.*, videos) into compact latent representations and learns the transition function in this latent space to forecast future states. By learning the transition function in this compressed representation rather than in pixel space, these models are trained to focus on features that are predictive of future latent states, such as motion and object dynamics, while ignoring irrelevant appearance details. This leads to representations that emphasize fundamental scene properties like structure, object permanence, and trajectory continuity [23]. As a result, they are ideally suited to serve as reward models for physics plausibility. VJEPA-2 [4] is one particular instantiation of a latent world model, which we employ in this work and which has demonstrated strong physics understanding. Yet, our approach of using latent world models as a source of reward is not tied to this particular instantiation; any suitable latent world model could potentially be used for this purpose.

In this paper, we formulate the problem of improving the physics correctness of video generation as an *inference-time alignment* problem [53, 57] that leverages a *latent world model* with high physics understanding as a reward model to search for physically plausible videos within the manifold learned by the generative model. In particular, we introduce `WMReward` by re-purposing VJEPA-2's surprise score as a reward function and show that it can be effectively used both as a Best-of-N (BoN) selector and as a guidance signal during generation to improve the physics of videos.

We validate the effectiveness of `WMReward` by achieving substantial gains in physics plausibility of video generation using three state-of-the-art video generative models: MAGI-1 [52], Sora2 [45], and a video latent diffusion model (vLDM). These models span diverse architectures, including large-scale autoregressive generation producing videos chunk-by-chunk and holistic diffusion-based generation. We obtain results across text-to-video (T2V), image-and-text-to-video (I2V), and video-and-text-to-video (V2V) settings. In particular, we achieve a new state of the

art on the challenging *PhysicsIQ* benchmark [44] with a final score of 62.0%, surpassing the previous best by 6.78%. This improvement is illustrated by the qualitative examples in Figure 2 and further validated by an additional 11.4% improvement over baselines in a human-preference study. Moreover, as shown in Figure 2, our VJEPA-based method demonstrates a strong scaling effect with the size of the search space, and outperforms VLM-based selectors (*e.g.*, Qwen3-VL [65]) that perform near chance level. With these results, we demonstrate the viability of using latent world models to improve the physical plausibility of video generation, paving the way toward developing reliable reward models for video generation inference-time alignment. The key contributions of our work are as follows: **(1)** We show that latent world models are useful to improve video generation. To this end, we devise an effective physics plausibility reward model, `WMReward`, instantiated with VJEPA's surprise score. **(2)** We highlight the scaling behavior of `WMReward` and show that model performance improves when increasing search space – with guided sampling showing more promising scaling. **(3)** We boost the physics plausibility in video generation across T2V, I2V and V2V settings and datasets, and achieve a final score of 62.0% on PhysicsIQ, outperforming prior state-of-the-art by 6.78%, further validated through a human-preference study showing a boost of 11.4% in win rate over the baseline.

2. Methodology: `WMReward`

In the following, we present `WMReward`, which transfers the physics prior from a latent world model (*e.g.*, VJEPA-2) into a physics plausibility reward signal. We leverage such reward signal to align the video generative model with the latent world model by sampling from a tilted distribution and thus generate physically more plausible videos.

2.1. Preliminaries

Diffusion and flow-matching models are two generative modeling paradigms that learn to transform Gaussian noise into samples from a target distribution $p_{\text{data}}(x)$. They do this by learning score functions $\nabla_{x_t} \log p_t(x_t)$ of the random variables x_t with distribution

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \tag{1}$$

where $x_0 \sim p_{\text{data}}$ and $\epsilon \sim \mathcal{N}(0, I)$. Here, α_t and σ_t are real-valued time-dependent functions where at $t = 0$, we have $\alpha_0 = 1$ and $\sigma_0 = 0$ so that $p_0(x) = p_{\text{data}}(x)$, and at $t = T$, $p_T(x)$ is approx. Gaussian. This score is learned by denoising score matching [58] or conditional flow matching[40].

Once this score is available, we can construct an SDE or ODE whose marginals $p_t(x)$ depend only on α_t, σ_t, and the learned score. We then can draw approximate samples from this learned distribution $p(x)$ by solving this SDE/ODE backward from $t = T$ to $t = 0$.

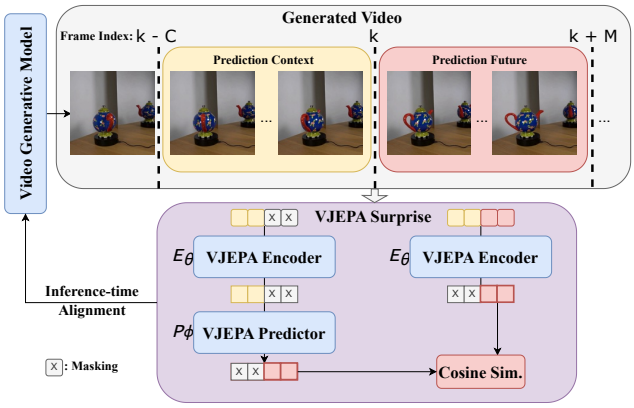


Figure 3. **Method Overview.** We leverage a latent world model, VJEPA-2, to steer video generative models for better physics plausibility. During generation, we apply a sliding window approach and split the generative model's output into sets of context and future frames. We encode generated context frames and predict the embedding of future frames using the latent world model's predictor. Then, we encode the generated future frames and compute the cosine similarity between its embedding and the latent world model prediction, referred to as surprise score. The surprise score serves as reward to search and guide the denoising trajectories.

2.2. Steering for Better Physics Plausibility

We frame improving physics plausibility in video generation as sampling from a reward-weighted tilted distribution:

$$p^*(x) \propto w(x)p(x), \tag{2}$$

where $p(x)$ is the pre-trained video model distribution and $w(x) > 0$ is a weighting function constructed from the reward $r(x)$ that can evaluate the physics plausibility.

To successfully draw samples from this tilted distribution, two key questions arise: 1) how to define the reward function that captures physics plausibility from a latent world model (Section 2.3); and 2) how to realize sampling from $p^*(x)$ given the pre-trained generative model $p(x)$ and the reward function $r(x)$ (Section 2.4). We depict an overview of our method in Figure 3.

2.3. Latent World Model Reward Signal

Recent work shows that latent world models—which learn to make predictions on compressed latent representations rather than raw pixels—can learn to judge physical plausibility [13, 23, 42]. By operating in a latent space, they tend to ignore superficial visual details and focus on underlying physical dynamics. One approach for training such models relies on self-supervised learning (SSL); VJEPA-2 [4] is an SSL-trained latent world model which achieves state-of-the-art performance in physics understanding [23], making it an ideal foundation for a reward model.

VJEPA learns representations through self-supervised

prediction of unmasked video regions from masked ones. The architecture comprises a context encoder E_θ that embeds video frames, and a predictor network P_ϕ that reconstructs target representations from partial observations. During training, a masked view x_{masked} is created from a video x by removing certain spatiotemporal regions. The predictor P_ϕ takes the encoder's output $E_\theta(x_{\text{masked}})$ along with learnable mask tokens Δ_m, indicating masked positions, and attempts to reconstruct the target representation. The target is computed using an exponentially-moving-averaged (EMA) encoder $\bar{E}_\theta$ applied to the full video. The training objective is:

$$\mathcal{L} = \|P_\phi(\Delta_m, E_\theta(x_{\text{masked}})) - \text{sg}(\bar{E}_\theta(x))\|_1 \tag{3}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation, preventing gradients from flowing through the target. The idea is that by making predictions in feature space, the model is encouraged to learn high-level spatiotemporal features that capture predictable dynamics rather than pixel-level details.

To turn VJEPA into a physics plausibility reward model, we leverage a simple intuition: being a world model, VJEPA should resonably predict the future in physically plausible videos. Hence, the more its predictions diverge from the generations, the more likely it is that the videos are less physically plausible.

We materialize this intuition by designing a reward function, `WMReward`, that measures prediction surprise by contrasting VJEPA-2 predictions with the generated video as shown in Figure 3. Concretely, we slide a window of length $C + M$ across a generated video x, where C denotes context frames and M denotes prediction horizon. At each position k, VJEPA observes context frames $x^{k-C+1:k}$ and must predict representations for future frames $x^{k+1:k+M}$. We produce latent world model's future representations using only context, as follows:

$$\hat{z}_k = P_\phi(\Delta_m, E_\theta(x^{k-C+1:k})), \tag{4}$$

where Δ_m indicates that M future positions are masked. We obtain generated video's representations by processing the complete window:

$$z_k = E_\theta(x^{k-C+1:k+M}). \tag{5}$$

Both z's contain representations for all $C+M$ positions; we extract and compare only the future portions corresponding to positions $C + 1$ through $C + M$ ($\hat{z}_k^{\text{fut}}$ and z_k^{fut}):

$$r(x) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left(1 - \cos(\hat{z}_k^{\text{fut}}, z_k^{\text{fut}})\right), \tag{6}$$

where $\mathcal{K}$ contains all valid windows. Videos receive higher rewards if their generated futures closely match VJEPA predictions, effectively measuring physical coherence.

$$\mathcal{L} = 1 - \cos(z_{pred},\ z_{gt}) \tag{7}$$

2.4. Instantiation of Sampling Schemes

We now consider practical sampling schemes with different weighting functions $w(x)$ to draw samples from the defined tilted distribution in Equation (2). As presented in Section 2.3, the reward function $r(x)$ is differentiable with respect to the video x, which allows us to leverage both gradient-based and gradient-free sampling schemes. We focus on three sampling schemes chosen for their representativeness, simplicity, and performance: guidance as an example for gradient-based method, best-of-N for gradient-free method, and their combination.

(i) **Guidance (∇)** uses $w(x) = \exp(\lambda r(x))$, where $r(x)$ is our reward function and $\lambda > 0$ is a temperature parameter controlling how much we upweight high-reward samples. Under the same noising process as in Equation (1), one can show that the score functions $\nabla_{x_t} \log p_t^*(x_t)$ of the time-dependent marginals $p_t^*(x_t)$ are given by

$$\nabla_{x_t} \log p_t(x_t) + \nabla_{x_t} \log \mathbb{E}\left[e^{\lambda r(x_0)} \,\middle|\, x_t\right]. \tag{8}$$

Therefore, approximating $\mathbb{E}[e^{\lambda r(x_0)} \mid x_t]$ as $e^{\lambda r(\mathbb{E}[x_0|x_t])}$, and using Tweedie's formula [22, 50],

$$x_{0|t} := \mathbb{E}\left[x_0 \mid x_t\right] = \frac{1}{\alpha_t}\left(x_t + \sigma_t \nabla_{x_t} \log p_t(x_t)\right), \tag{9}$$

we obtain the following approximation to Equation (8)

$$\nabla_{x_t} \log p_t^*(x_t) \approx \nabla_{x_t} \log p_t(x_t) + \lambda \nabla_{x_t} r(x_{0|t}(x_t)) \tag{10}$$

We can therefore sample from $p^*(x)$ by using this new score approximation in the SDE/ODE sampler.

(ii) **Best-of-N Search** (BoN) involves generating N independent samples $\{x^{(i)}\}_{i=1}^N$ (*i.e.*, particles) from the base model $p(x)$ and selecting the sample with highest reward:

$$x^* = \text{argmax}_{x \in \{x^{(i)}\}_{i=1}^N} r(x). \tag{11}$$

This procedure effectively samples from a distribution

$$p^*(x) \propto p(x)[F(r(x))]^{N-1}, \tag{12}$$

where $F(\cdot)$ is the cumulative distribution function (CDF) of $r(x)$ under $p(x)$ and $w(x) = [F(r(x))]^{N-1}$.

(iii) **∇ + BoN** combines both approaches by using the guidance scheme to generate N samples and then chooses the highest reward sample among them. This yields approximate samples from the tilted distribution with weight function $w(x) = \exp(\lambda r(x)) \cdot [F_\lambda(r(x))]^{N-1}$, where $F_\lambda(\cdot)$ is the CDF of $r(x)$ under the guided distribution. This combination is beneficial because it achieves stronger tilting without requiring large λ values (where the score approximation becomes inaccurate), and best-of-N selection provides an additional strategy for filtering samples where the approximation might have been sub-optimal. In this way, we benefit from the increased likelihood of high-reward samples provided by guidance, while mitigating approximation errors.

3. Experiments

We show the effectiveness of sampling with `WMReward` for improving physics plausibility across three video generation setups: image-and-text-to-video (I2V), video-and-text-to-video (V2V), and text-to-video (T2V). For all experiments, we apply `WMReward` to MAGI-1 24B [52] and to vLDM 5B*, and additionally to Sora2 [45] on PhysicsIQ.

3.1. Image and Multiframe-conditioned Generation

We start by evaluating `WMReward` on PhysicsIQ [44], a benchmark for I2V and V2V generation. We follow their setup and use the text prompt along with a three-second context video (for V2V) or the last frame (for I2V). Given the conditioning, we generate five-second video continuations, evaluate it against the ground-truth and compute the PhysicsIQ score that relies on an aggregation of four metrics: spatial intersection over union (IoU), spatio-temporal IoU over motion masks, a weighted spatial IoU, and pixel MSE. For all methods, we use 16 particles for the search.

Table 1 demonstrates that sampling with `WMReward` consistently outperforms vanilla sampling for vLDM, MAGI-1, and Sora2, achieving state-of-the-art performance (best and second-best) across baselines and evaluation dimensions. In particular, by using the appropriate sampling strategy, ∇+`BoN`, we achieve 62.0% final PhysicsIQ score on V2V generation, which significantly outperforms the previous state-of-the-art MAGI-1 model by 6.78%. For I2V, sampling with `WMReward` surpasses the previous state-of-the-art Sora2 [45] by 4.13% on PhysicsIQ.

Comparing Reward Signals. We compare `WMReward` against alternative physics plausibility signals based on foundation models such as VideoMAE and Qwen VL (2.5-7B-Instruct and 3-8B-Instruct). For VideoMAE [56], a self-supervised video masked autoencoder that learns spatiotemporal representations by reconstructing masked patches in pixel space, we use a surprise score given by the reconstruction error between the predicted pixels and the generated pixels as reward signal [23]. For Qwen2.5-VL [5] and Qwen3-VL [65], following Jang et al. [29], we pose a physics-plausibility question and request a binary 0/1 answer, then follow practice in [39] and use the corresponding positive token logit as the reward.

As shown in Table 1 and Figure 2, `WMReward` yields stronger performance than alternative signals, indicating that predictive latent-space surprise is a more effective proxy for physics plausibility than pixel reconstruction or VLM judgment, which aligns with recent work suggesting that latent world models exhibit better physics understanding [23]. With these results, we show that the physics knowledge in latent world models may be transferred as reward signal to improve video generation.

Table 1. **Image and Multiframe-conditioned Generation Performance on PhysicsIQ.** Our results are highlighted in gray and performance changes against baseline sampling are indicated in green. Best results are highlighted in **bold** and second-best are underlined. For all search methods, we use 16 particles.†

	Spatial IoU ↑	Spatio-temp. IoU ↑	Weighted Spatial IoU ↑	MSE ↓	PhysicsIQ Score ↑
I2V Generation					
Sora [14]	0.138	0.047	0.063	0.030	10.00
Pika 1.0 [47]	0.140	0.041	0.078	0.014	13.00
SVD [12]	0.132	0.076	0.073	0.021	14.80
Lumiere [9]	0.113	0.173	0.061	0.016	19.00
VideoPoet [31]	0.141	0.126	0.087	0.012	20.30
Wan2.1 [59]	0.153	0.100	0.112	0.023	20.89
Gen 3 [51]	0.201	0.115	0.116	0.015	22.80
Kling1.6 [33]	0.197	0.086	0.144	0.025	23.64
VLIPP [67]	N/A	N/A	N/A	N/A	34.6
vLDM	0.221	0.120	0.144	0.008	27.76
+ `VideoMAE(BoN)`	0.229	0.132	0.152	0.009	29.42 (+1.66)
+ `Qwen2.5-VL(BoN)`	0.228	0.092	0.143	0.009	26.21 (-1.55)
+ `Qwen3-VL(BoN)`	0.214	0.139	0.141	0.008	28.51 (+0.75)
+ `WMReward(`∇`)`	0.221	0.122	0.144	0.009	27.88 (+0.12)
+ `WMReward(BoN)`	0.249	0.155	0.170	0.008	32.90 (+5.14)
+ `WMReward(`∇`+BoN)`	0.234	0.175	0.165	**0.007**	33.44 (+5.68)
MAGI-1 [52]	0.252	0.146	0.151	0.011	29.77
+ `VideoMAE(BoN)`	0.251	0.151	0.153	0.012	29.95 (+0.18)
+ `Qwen2.5-VL(BoN)`	0.241	0.093	0.141	0.013	24.99 (-4.78)
+ `Qwen3-VL(BoN)`	0.247	0.161	0.146	0.010	30.21 (+0.44)
+ `WMReward(`∇`)`	0.252	0.145	0.152	0.010	29.77 (+0.00)
+ `WMReward(BoN)`	0.253	**0.239**	0.162	0.008	36.56 (+6.79)
+ `WMReward(`∇`+BoN)`	0.267	0.218	0.168	0.008	36.28 (+6.51)
Wan2.2 [59]	0.358	0.120	0.217	0.008	38.26
+ `VideoMAE(BoN)`	0.347	0.124	0.216	0.008	37.91 (-0.35)
+ `Qwen2.5-VL(BoN)`	0.356	0.116	0.213	0.009	37.58 (-0.68)
+ `Qwen3-VL(BoN)`	0.353	0.128	0.216	0.008	38.51 (+0.25)
+ `WMReward(BoN)`	**0.378**	0.171	**0.247**	0.008	44.39 (+6.13)
Sora2 [45]	0.363	0.187	0.231	**0.007**	42.30
+ `VideoMAE(BoN)`	0.365	0.193	0.234	**0.007**	42.94 (+0.64)
+ `Qwen2.5-VL(BoN)`	0.340	0.166	0.210	**0.007**	38.69 (-3.61)
+ `Qwen3-VL(BoN)`	0.349	0.206	0.226	**0.007**	42.54 (+0.24)
+ `WMReward(BoN)`	0.377	0.229	0.246	**0.007**	**46.43** (+4.13)
V2V Generation					
Lumiere [9]	0.170	0.155	0.093	0.013	23.00
VideoPoet [31]	0.204	0.164	0.137	0.010	29.50
MAGI-1 [52]	0.416	0.279	0.294	**0.005**	55.22
+ `VideoMAE(BoN)`	0.407	0.286	0.285	**0.005**	54.71 (-0.51)
+ `Qwen2.5-VL(BoN)`	0.415	0.272	0.293	**0.005**	54.71 (-0.51)
+ `Qwen3-VL(BoN)`	0.407	0.279	0.289	**0.005**	54.42 (-0.80)
+ `WMReward(`∇`)`	0.414	0.279	0.292	**0.005**	55.02 (-0.20)
+ `WMReward(BoN)`	0.435	0.324	0.316	**0.005**	60.34 (+5.12)
+ `WMReward(`∇`+BoN)`	**0.439**	**0.339**	**0.325**	**0.005**	**62.00** (+6.78)

Human Study. We also supplement our evaluations with a human study to verify the effectiveness of `WMReward`. We conduct a human study on the full PhysicsIQ benchmark with a side-by-side comparison interface where five annotators view pairs of generated videos along with the original text prompt and additional conditioning information. For each video pair, annotators provide judgments across three

*vLDM is a latent video diffusion model that utilizes a DiT backbone and jointly models spatial and temporal components at once.

†ICCV'25 PhysicsIQ Challenge's official results: MAGI-1 I2V 37.39 (+7.62) & V2V 62.64 (+7.42). Sora2 is an API model; ∇ not available.

Table 2. **Human evaluation results on PhysicsIQ and VideoPhy.** Pairwise human preference judgments across three criteria: Physics Plausibility, Visual Quality, and Prompt Alignment. Winning rate (Win) is the fraction of non-neutral comparisons won, $\frac{\text{wins}}{\text{wins}+\text{losses}} \times 100$. Accuracy is computed as $\frac{\text{wins}+0.5\times\text{neutrals}}{\text{total}} \times 100$ to account for tie. Higher is better, best in bold.

	Physics Plausibility		Visual Quality		Prompt Alignment		Overall	
	Win	Acc.	Win	Acc.	Win	Acc.	Win	Acc.
PhysicsIQ								
vLDM	46.9	45.2	45.3	41.2	48.7	46.8	47.0	44.3
+ `WMReward(∇+BoN)`	**53.1**	**54.8**	**54.7**	**58.8**	**51.3**	**53.2**	**53.0**	**55.7**
MAGI-1	45.1	41.3	47.1	42.9	44.3	35.1	45.5	40.0
+ `WMReward(∇+BoN)`	**54.9**	**58.7**	**52.9**	**57.1**	**55.7**	**64.9**	**54.5**	**60.0**
VideoPhy								
vLDM	43.8	38.0	47.4	41.9	**51.0**	**53.3**	47.4	43.2
+ `WMReward(∇+BoN)`	**56.2**	**62.0**	**52.6**	**58.1**	49.0	46.7	**52.6**	**56.8**
MAGI-1	40.7	30.4	42.8	33.3	49.5	48.6	44.3	36.8
+ `WMReward(∇+BoN)`	**59.3**	**69.6**	**57.2**	**66.7**	**50.5**	**51.4**	**55.7**	**63.2**

Table 3. **Text-conditioned Generation Performance on VideoPhy.** Physics Consistency (PC) and Semantic Adherence (SA) are computed with a VLM-based evaluator that measures how well generated videos follow physics laws and text prompts. Higher is better, best is bold. For all search methods, we use 8 particles.

	Solid-Solid		Solid-Fluid		Fluid-Fluid		Overall	
	SA	PC	SA	PC	SA	PC	SA	PC
VideoCrafter2 [18]	50.4	32.2	50.7	27.4	48.1	29.1	50.3	29.7
DreamMachine [41]	55.1	21.7	59.6	23.3	58.2	18.2	57.5	21.8
LaVIE [62]	40.8	18.3	48.6	37.0	69.1	50.9	48.7	31.5
HunyuanVideo [32]	55.2	16.1	67.1	30.1	54.5	54.5	60.2	28.2
vLDM	46.6	20.7	67.5	28.6	52.4	48.6	56.9	28.0
+ `WMReward(BoN)`	42.7	**28.7**	59.6	36.3	38.5	**50.0**	50.3	**34.9**
+ `WMReward(∇+BoN)`	39.2	25.2	65.1	**37.0**	50.0	**50.0**	53.5	34.3
Wan2.2 [59]	52.7	13.9	71.9	**17.8**	67.5	**30.2**	63.5	18.1
+ `WMReward(BoN)`	43.7	**21.8**	63.8	15.9	53.8	26.9	55.2	**20.9**
MAGI-1 [52]	42.2	19.0	67.2	27.7	51.9	33.5	54.4	25.0
+ `WMReward(BoN)`	30.8	28.0	59.6	**36.3**	45.5	36.4	45.3	32.8
+ `WMReward(∇+BoN)`	29.4	**28.7**	61.6	30.1	40.0	**52.7**	44.8	**33.1**

criteria: Physics Plausibility, Visual Quality, and Prompt Alignment (further details in Appendix Section C). We obtain two groups of 198 annotations for both vLDM (I2V) and MAGI-1 (V2V) generations, comparing the baseline vanilla sampling strategy against `WMReward(∇+BoN)`. For each criterion, the annotators select their preference among the shown videos or mark them as neutral. Results are aggregated using win rates (excluding neutrals) and accuracy scores to account for ties ((wins + 0.5 × neutrals) / total). Table 2 demonstrates that `WMReward` delivers significant improvement in all three evaluation criteria, with the winning rate in physics plausibility being most remarkable.

3.2. Text-conditioned Generation

For evaluating the T2V setup, we rely on VideoPhy [7]. Following practice in Zhang et al. [73] and Wang et al. [60], we adopt the automatic VLM-based evaluator, which queries a VLM with templated questions to score semantic adherence (SA) and physics consistency (PC) on a scale of 1-5 for each generated video. We report the per-axis pass rates with each metric larger than 4 over the 344 benchmark prompts. For all methods, we use 8 particles for search.

Table 3 shows that incorporating `WMReward` into sampling substantially improves the physics consistency of both MAGI-1 and vLDM by 8.1% and 6.9%, respectively, surpassing all baseline models. We also note that the semantic adherence decreases potentially due to the VJEPA surprise not including semantic information from the text-condition, and to further verify the trade-off between semantic adherence and physics consistency we perform a human study.

Human Study. To complement the VLM-based evaluation, we conduct a human study following Section 3.1's protocol for a subset of 100 VideoPhy generations. Human judgment, see Table 2, highlights improvements in both physics plausibility and visual quality when using `WMReward` for T2V generation. The slight decrease in prompt alignment for vLDM is consistent with the text-agnostic nature of VJEPA's surprise reward, but the drop is small relative to the gains, yielding a net improvement in the overall metric. We expect this limitation can be mitigated by developing compositional or text-conditioned physics plausibility rewards, an interesting direction for future work.

3.3. Scaling the Search Space of `WMReward`

We now study the scaling behavior of `WMReward` with the size of the search space. In particular, we scale the number of particles and display results in Figure 2 (right). We observe that, by scaling up the number of particles, there is a steady improvement of PhysicsIQ score. The performance boost is most significant for $N \leq 4$, and steadily increase for larger N. Additionally, as the number of particles increases, the performance variance is smaller and gradually stabilizes. Moreover, as shown in Figure 4, allocating more compute by increasing the number of particles and applying guidance steers the PhysicsIQ score distribution to concentrate in the high-score region. The shift of distributions from $N = 1$ to $N = 16$ shows the effectiveness of `WMReward(BoN)` by using more particles, while `WMReward(∇+BoN)` with guidance further sharpening the upper tail of the distribution, confirming its stronger scaling behavior. The consistent scaling behavior of `WMReward` on PhysicsIQ suggests that by increasing inference compute allocation, we improve the physical plausibility of video generation.

Runtime and Memory. Inference-time alignment methods trade additional computation for improved

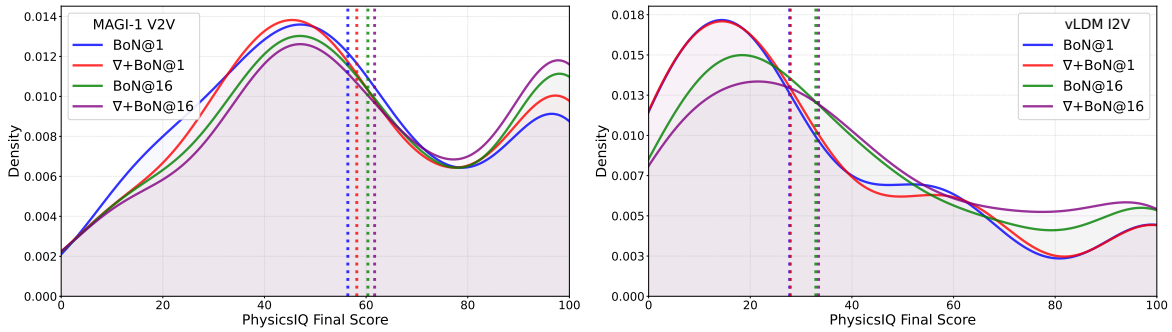


Figure 4. **Improving Performance via Particle Scaling and Guidance.** PhysicsIQ score distributions with Gaussian KDE for MAGI-1 V2V (left) and vLDM I2V (right). Scaling the number of particles (here, 1 vs. 16) yields substantial gains for `BoN`; adding guidance further sharpens the distribution toward higher physics plausibility. This leads to overall higher average score (dashed vertical line).

Table 4. **Computation Cost.** For each model, the first row shows absolute values (runtime in seconds, memory in GB per GPU). Subsequent rows show overhead multipliers relative to baseline. N is the search budget (number of particles). All experiments on H200 GPUs; MAGI-1 uses 8 GPUs, vLDM uses 1 GPU.

Method	Time	Memory	N
vLDM (baseline)	106.77 ± 0.43 s	27.64 GB	1
+ `WMReward(BoN)`	$\times N$	$\times$1.00	8
+ `WMReward(∇)`	$\times$5.02	$\times$4.27	1
+ `WMReward(∇+BoN)`	$\times 5.02N$	$\times$4.27	8
MAGI-1 (baseline)	265.66 ± 4.49 s	50.03 GB	1
+ `WMReward(BoN)`	$\times N$	$\times$1.00	16
+ `WMReward(∇)`	$\times$4.96	$\times$2.07	1
+ `WMReward(∇+BoN)`	$\times 4.96N$	$\times$2.07	16

performance—a recent paradigm shift, from long reasoning traces to test-time search in general. In Table 4, we show the computational overhead induced by `WMReward`. `WMReward(BoN)` is parallelized as N vanilla sampling trajectories with the same memory bound, and its memory footprint scales linearly with the number of particles. Guidance introduces extra computation mainly from the gradient backpropagation, and is proportional to the number of steps where guidance is performed through the denoising process. One can choose whether to incorporate first-order information of reward signal as budget allows. In general, `WMReward` can be adapted to the available computation budget with appropriate sampling methods, offering a spectrum of compute-performance tradeoffs.

3.4. Analysis and Ablation

Having established that `WMReward` achieves strong improvements in physics plausibility and scales effectively with computational budget, we now investigate several key questions about the method's properties.

Does `WMReward` impair perceptual qualities? To assess whether physics improvements come at the cost of general perceptual quality, we use the VBench evaluators [26] to rate generated PhysicsIQ and VideoPhy videos

Table 5. **General Visual Quality with VBench Evaluators.** Visual quality metrics computed on videos generated for the PhysicsIQ and VideoPhy datasets. Higher is better for all metrics.

	Subject Consistency	Background Consistency	Motion Smoothness	Temporal Flickering	Imaging Quality	Aesthetic Quality
			PhysicsIQ			
vLDM	94.68	95.60	99.57	99.48	67.47	48.94
+ `WMReward(∇)`	94.64	95.58	99.56	99.47	67.50	48.95
+ `WMReward(BoN)`	95.29	95.89	99.62	99.61	67.71	49.19
+ `WMReward(∇+BoN)`	95.25	95.89	99.62	99.61	67.83	49.03
MAGI-1	95.82	97.25	99.70	99.82	69.07	47.05
+ `WMReward(∇)`	95.90	97.25	99.70	99.82	69.15	47.05
+ `WMReward(BoN)`	95.87	97.30	99.71	99.84	69.24	46.88
+ `WMReward(∇+BoN)`	95.88	97.30	99.71	99.85	69.25	46.94
			VideoPhy			
vLDM	93.73	95.79	98.57	97.80	60.34	49.71
+ `WMReward(∇)`	93.72	95.78	98.57	97.80	60.05	49.69
+ `WMReward(BoN)`	94.82	96.31	98.69	98.02	59.44	49.54
+ `WMReward(∇+BoN)`	94.67	96.36	98.72	98.13	60.03	49.87
MAGI-1	94.43	95.87	98.80	98.08	56.15	46.39
+ `WMReward(∇)`	94.58	95.92	98.81	98.12	56.17	46.46
+ `WMReward(BoN)`	96.27	96.68	98.92	98.44	58.15	47.54
+ `WMReward(∇+BoN)`	96.54	96.87	98.97	98.54	57.84	47.73

across six key visual quality metrics: subject consistency, background consistency, motion smoothness, temporal flickering, imaging quality, and aesthetic quality. As shown in Table 5, image quality and aesthetic quality tend to show small improvements, possibly due to the stronger physics plausibility achieved by `WMReward`. This trend aligns with the human-preference results in Table 2. We also observe improved temporal consistency, motion smoothness, and temporal flickering, suggesting that by following physics principles, such as artifact suppression, and adherence to mass conservation and continuity, the overall perceptual quality of generated videos are also improved.

How does `WMReward` perform with other sampling methods? In principle, any valid sampling scheme can be used to sample from the `WMReward` tilted distribution. We further demonstrate the compatibility of `WMReward` with alternative sampling schemes: SMC [53] and SVDD [36] (detailed in Section 4), using vLDM on PhysicsIQ. Empirically, both SMC and SVDD deliver gains over vanilla

Table 6. **Comparison of Sampling Methods.** vLDM I2V PhysicsIQ generation under different sampling schemes. Columns indicate number of particles (N), showing scaling effects.

	2	4	8	16
`WMReward(SMC)`	27.14	26.96	28.12	29.24
`WMReward(SVDD)`	27.31	28.57	28.54	28.87
`WMReward(BoN)`	27.76	**31.87**	32.32	32.90
`WMReward(`∇`+BoN)`	**27.88**	31.05	**32.62**	**33.44**

sampling, yet, they still underperform `BoN` and ∇+`BoN`, as shown in Table 6. Intuitively, since VJEPA-2 surprise is differentiable, its gradients provide step-wise guidance that plays a role similar to, and often more accurate than, the multi-rollout importance estimates by SVDD. Also, SMC shows smaller gains, likely because early-stage reward estimates are noisy, causing resampling to collapse on suboptimal seeds. Practically, SMC also requires multiple parallel denoising trajectories per sample for resampling, which increases the memory footprint and is nontrivial to parallelize. For these reasons, we adopt `WMReward(`∇`+BoN)` as a more practically effective sampling scheme.

4. Related Work

Video Diffusion Models. Diffusion and flow-based video generative models [9, 14, 31, 45, 52, 59] learn to generate video sequences by reversing a noising process on temporal visual data. While they achieve promising visual quality in generation, they still often produce physically implausible videos [7, 8, 44, 69]. To mitigate this shortcoming, there has been substantial work devising pre- and post-training approaches focused on explicitly injecting physics constraints [70], enabling motion prediction [17], enhancing dynamics [38] or leveraging learned physics priors from foundation models through finetuning or distillation [16, 71, 73]. Yet, another underexplored line of work has focused on devising inference-time methods to improve physics. Particularly, Xue et al. [64] explore the learned manifold of the generative model through prompt rewriting and Yang et al. [67] perform motion planning with VLMs on top of a motion-controllable video generative model [15].

Physics Understanding in Vision Models. Physics understanding is essential to a model's ability to reason and predict scenes under physical laws [11, 13, 49], and in particular, it is important to generate physically plausibility videos [44, 69]. Recent work [13] has investigated intuitive physics understanding of vision models, including VLMs [5, 19], generative models [1], and self-supervised learning models [4, 56, 61], showing that video SSL approaches based on the Joint-Embedding-Predictive-Architecture (JEPA) [4, 10] exhibit the highest performance among predictive methods. Moreover, intuitive physics properties like object permanence and shape consistency have been shown to emerge in video-JEPA (VJEPA) models [23]. VJEPA consists of a video encoder and a predictor trained on internet-scale data with a masked predictive objective in the representation space. Predicting the future in the representation space reduces the sensitivity to appearance details and holds the potential to emphasize other aspects like dynamics and interaction.

Inference-Time Alignment for Image Generation Models. Inference-time alignment steers a pre-trained diffusion model toward desired properties during generation without retraining. For one, there are derivative-free approaches [28, 35–37, 46, 53, 63]. Among those, BoN selects the highest-scoring candidates at the end of the denoising process [43]. SMC [53, 63] assigns importance weights to multiple denoising trajectories and perform resampling during generation. Value-based importance sampling methods (SVDD [36]) steer a single trajectory by resampling according to the local step reward. On the other hand, we have gradient-based approaches that adjust the score function [2, 3, 6, 20, 24, 55, 68], steering the denoising process toward high reward regions– resulting in *e.g.*, high utility or high diversity samples. These assume access to differentiable reward functions. Despite their different implementations, these methods share a similar principle of using reward signals to refine local steps or enable better global exploration.

5. Conclusion and Discussion

In this work, we presented `WMReward`, an inference-time alignment method to improve physics plausibility of video generative models, and highlighted the viability of using latent world models (*i.e.*, VJEPA-2) as reward signal. Throughout our experimentation, we demonstrated the potential of leveraging latent world model's to improve physics in video generation, yielding substantial improvements on the challenging PhysicsIQ and VideoPhy benchmarks without requiring any further training.

We see three potential avenues worth exploring: (1) *Improving video generative models*: The quality of video generative model determines the set of potential solutions that can be explored with our method; developing stronger, more searchable models is very promising. (2) *Improving reward models:* The strength of the reward model heavily influences the overall performance of the system. Although VJEPA-2 is already informative, its training data is currently limited and as a result the model does not cover all physics phenomena, *e.g.*, we observe that material understanding like weight and friction is limited. Future work on latent world models could provide broader understanding of physical phenomena. (3) *Improving search algorithms:* Better and more efficient search algorithms offer a promising avenue to improve performance. Early during diffusion, the intermediate estimated video samples $x_{0|t}$ are blurry, potentially leading to unreliable reward predictions.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 8, 13

[2] Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *European Conference on Computer Vision*, pages 213–229. Springer, 2024. 2, 8

[3] Reyhane Askari-Hemmat, Mohammad Pezeshki, Elvis Dohmatob, Florian Bordes, Pietro Astolfi, Melissa Hall, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Improving the scaling laws of synthetic data with deliberate practice. In *ICML*, 2025. 2, 8

[4] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. 2, 3, 8

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 8, 14

[6] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 843–852, 2023. 8

[7] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 2, 6, 8

[8] 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. *arXiv preprint arXiv:2503.06800*, 2025. 2, 8

[9] 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, 2024. 1, 5, 8

[10] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. 8

[11] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. 8

[12] 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, 2023. 5

[13] Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. 2, 3, 8

[14] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 1, 5, 8

[15] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025. 2, 8

[16] Qinglong Cao, Ding Wang, Xirui Li, Yuntian Chen, Chao Ma, and Xiaokang Yang. Teaching video diffusion model with latent physical phenomenon knowledge. *arXiv preprint arXiv:2411.11343*, 2024. 2, 8

[17] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In *Forty-second International Conference on Machine Learning*, 2025. 2, 8

[18] 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, 2024. 6

[19] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 8

[20] Nicola Dall'Asen, Xiaofeng Zhang, Reyhane Askari Hemmat, Melissa Hall, Jakob Verbeek, Adriana Romero-Soriano, and Michal Drozdzal. Increasing the utility of synthetic images through chamfer guidance. *arXiv preprint arXiv:2508.10631*, 2025. 2, 8

[21] Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. Prompt expansion for adaptive text-to-image generation. In *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. 2

[22] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 4

[23] 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. 2, 3, 5, 8, 14

[24] Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023. 2, 8

[25] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[26] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7

[27] Tariq Berrada Ifriqi, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models, 2025. 2

[28] Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. *arXiv preprint arXiv:2506.20701*, 2025. 8

[29] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025. 5

[30] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 2

[31] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. 1, 5, 8

[32] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 6

[33] Kuaishou. Kling video model, 2024. 5

[34] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 2

[35] Gyubin Lee, Truong Nhat Nguyen Bao, Jaesik Yoon, Dongwoo Lee, Minsu Kim, Yoshua Bengio, and Sungjin Ahn. Adaptive inference-time scaling via cyclic diffusion search. *arXiv preprint arXiv:2505.14036*, 2025. 8

[36] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, et al. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024. 7, 8

[37] Xiner Li, Masatoshi Uehara, Xingyu Su, Gabriele Scalia, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Shuiwang Ji. Dynamic search for inference-time alignment in diffusion models. *arXiv preprint arXiv:2503.02039*, 2025. 2, 8

[38] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24142–24153, 2024. 2, 8

[39] 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*, pages 366–384. Springer, 2024. 5

[40] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[41] Luma AI. Dream machine | ai video generator. https://lumalabs.ai/dream-machine, 2024. 6

[42] Ge Ya Luo, Gian Mario Favero, ZhiHao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond FVD: An enhanced evaluation metrics for video generation distribution quality. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3

[43] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. 2, 8

[44] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. 2, 3, 5, 8

[45] OpenAI. Sora 2 system card. https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf, 2025. Accessed: January 21, 2026. 1, 2, 5, 8

[46] Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27326–27337, 2025. 8

[47] Pika Labs. Pika 1.0, 2023. 5

[48] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1

[49] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel

Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. 8

[50] Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992. 4

[51] Runway. Gen-3 Alpha, 2024. 5

[52] Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale, 2025. 1, 2, 5, 6, 8, 13

[53] 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. *arXiv preprint arXiv:2501.06848*, 2025. 2, 7, 8

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 13

[55] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 8

[56] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 5, 8

[57] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025. 2

[58] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. 3

[59] 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. 1, 5, 6, 8, 13

[60] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025. 6

[61] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 8

[62] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025. 6

[63] Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023. 8

[64] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18826–18836, 2025. 2, 8

[65] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3, 5, 14

[66] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 2

[67] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. *arXiv preprint arXiv:2503.23368*, 2025. 2, 5, 8

[68] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Y Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37: 22370–22417, 2024. 2, 8

[69] 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. *arXiv preprint arXiv:2510.11512*, 2025. 2, 8

[70] 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. *arXiv preprint arXiv:2509.21309*, 2025. 2, 8

[71] Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025. 2, 8

[72] Xiaofeng Zhang, Aaron Courville, Michal Drozdzal, and Adriana Romero-Soriano. The intricate dance of prompt complexity, quality, diversity, and consistency in t2i models, 2025. 2

[73] 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. *arXiv preprint arXiv:2505.23656*, 2025. 2, 6, 8

[74] Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. *arXiv preprint arXiv:2505.23614*, 2025. 2