Proprio: Latent Self-Scoring and Inference-Time Refinement for Physically Plausible Video Generation

Mariam Hassan [1], **Kaouther Messaoud** [2], **Wuyang Li** [1], **Alexandre Alahi** [1]
[1] École Polytechnique Fédérale de Lausanne (EPFL), [2] Télécom Paris, IP Paris
Project Page: `https://vita-epfl.github.io/Proprio/`

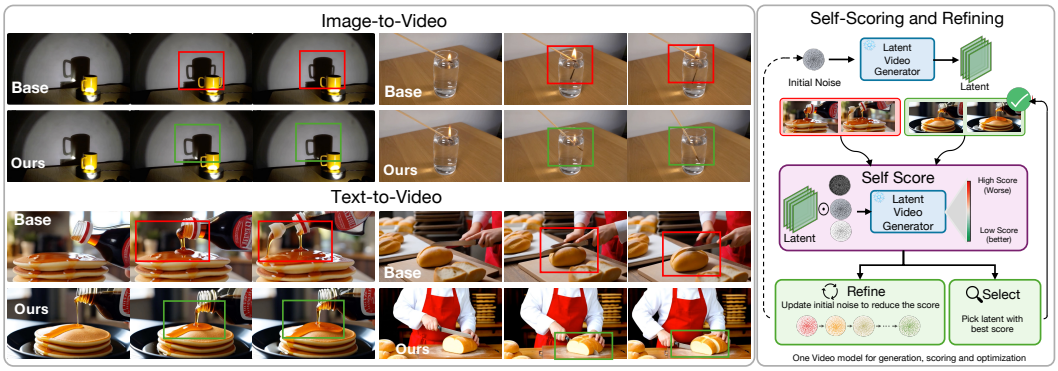


Figure 1: **Proprio enables a video generator to evaluate and improve its own generation.** Left: qualitative refinement comparisons for TurboWan2.2. Right: Proprio computes a latent self-score from the model's denoising residual and uses it for either sample selection or inference-time refinement.

Abstract

Modern video generative models produce visually impressive results, yet frequently violate basic physical principles. We propose **Proprio**, a training-free framework that enables a frozen video generator to assess and improve the physical plausibility of its own outputs. Inspired by *proprioception*, the biological sense of one's own movement, Proprio treats the model's flow residual under controlled latent perturbations as a self-scoring signal. Samples that are better explained by the generator's learned dynamics induce smaller and more stable residuals. We aggregate this signal across timesteps and perturbations, focus it on motion-relevant regions with a dynamic spatiotemporal mask, and use it for best-of-N search, gradient-based self-refinement, or both. Across text-to-video and image-to-video benchmarks, Proprio consistently improves physical plausibility, outperforming VLM-based scoring, and external world-model baselines in several settings. With TurboWan2.2, Proprio improves Physics-IQ from 32.2 to 37.5 (+16.5%) and VideoPhy2-hard physical commonsense from 45.6 to 55.0 (+20.6%). Human evaluation further shows that raters prefer Proprio-selected or refined videos for physical plausibility in roughly two-thirds of comparisons. These results suggest that frozen video generators contain actionable internal signals for evaluating and improving the physical plausibility of their own outputs.

1 Introduction

Diffusion models [13, 24] have achieved remarkable progress, emerging as dominant paradigms for high-fidelity video generation [7, 30, 34]. Despite their impressive visual quality, *these models still struggle to maintain physical plausibility*, often producing videos that violate basic real-world con-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

straints. Such failures not only limit perceptual realism but also undermine the utility of these models as reliable world models for downstream applications such as robotics, planning, and simulation.

To address this, existing efforts typically rely on external sources of feedback to provide physical assessment and supervision. Some work inject physical structure through curated datasets, physical priors, or differentiable simulation [35, 23]. Another line of work guides generation using external evaluators, including vision-language models, or pretrained latent world models such as V-JEPA [38, 8, 44, 42, 2, 6]. These methods provide useful signals for selecting physically plausible generations and have shown promising results in improving physical realism.

While effective, this external-feedback paradigm faces two key limitations. *(1) External evaluators introduce their own biases and failures*: VLMs struggle to assess physical plausibility because their judgments are often entangled with text-image alignment objectives and visual appearance [28]. Pretrained world models provide a more dynamics-oriented signal, but operate on decoded videos and map them into their own representation space, creating a representation mismatch with the generator's latent space. *(2) Misalignment with the generator's internal dynamics*: external evaluators are trained with different training objectives, so their feedback is not directly aligned with the generator's sampling process and may miss generator-specific inconsistencies across denoising steps.

These limitations motivate an intrinsic perspective on physical plausibility: *Can a video generator leverage its own denoising dynamics as a self-contained signal for producing more physically plausible videos?* Here, denoising dynamics refers to the model-predicted update direction at each noise level. We focus on its prediction error, which we call the *flow residual*: the discrepancy between the predicted denoising or velocity direction and the target induced by a controlled perturbation.

Conceptually, the flow residual can be interpreted as an approximate likelihood signal [19], providing inherent cues for assessing physical validity, as previously explored in curated simulated video pairs [41]. In contrast, we reinterpret this likelihood-preference signal as a generator-native objective. Since video generators are trained on large-scale real video datasets, their learned distributions implicitly capture regularities of real-world dynamics. It is therefore natural to treat consistency with such learned dynamics as an intrinsic proxy for physical plausibility. **Empirically**, we test this intuition with a small-scale diagnostic check. Given a set of real videos, we pair each video with a generated counterpart that exhibits a physical failure using a video model, and compare their residual scores under that model (see Appendix A.1). In 82.2% of the pairs, the residual signal prefers the ground-truth video over the model's own physically implausible generation, suggesting the model's internal denoising dynamics can distinguish natural video dynamics from its own failure cases.

Motivated by this, we propose ***Proprio***, a generator-native, training-free framework that converts internal denoising signals into an internal proxy for physical plausibility. Inspired by ***proprioception***, the biological ability to sense one's motion without external observation, Proprio enables a frozen video generator to assess and improve its own outputs. Specifically, we define a *noise-conditioned self-consistency signal* based on the model's denoising residual under controlled latent perturbations. We robustly estimate the score by aggregating residuals across multiple timesteps with *inverse-variance weighting*, which downweights noisy estimates, and by applying a *dynamic spatiotemporal mask* to emphasize motion-relevant regions where physical failures occur. Beyond scoring, Proprio uses the score as a self-refinement objective. It refines generations by backpropagating through the frozen sampler and updating the initial noise to produce a more internally consistent video.

We evaluate Proprio in three inference-time strategies: best-of-N search, self-refinement, and hybrid search-and-refine approach (see Fig. 1). For image-to-video (I2V) and text-to-video (T2V) generation, Proprio improves physical accuracy over VLM-based, and outperforms external pretrained world-model in several settings. These results suggest that a frozen video generator can be used to generate videos and to guide selection and refinement of its outputs. Our contributions are as follows:

- **Generator-native self-scoring:** We introduce Proprio, a training-free framework that reuses a frozen generator's denoising residual as an intrinsic score for physical plausibility.
- **Inference-time self-alignment:** We turn this score into an optimization objective for a frozen generator, refining its outputs by updating only the initial noise, and optionally combine this refinement with best-of-N selection.
- **Empirical validation:** We demonstrate consistent gains on I2V and T2V physics benchmarks, showing that a single frozen generator can act as both sampler and self-evaluator, guiding selection and refinement toward more physically plausible dynamics.

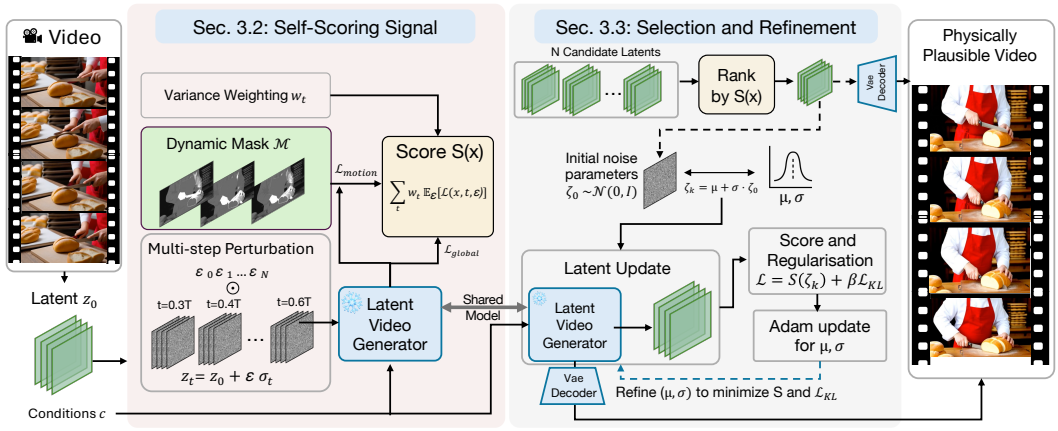


Figure 2: **Proprio method overview**. For a generated latent video, Proprio computes a per-video self-scoring signal by perturbing the latent at multiple timesteps, measuring denoising residuals with the frozen generator, weighting timesteps by inverse variance, and emphasizing informative motion regions with a dynamic spatiotemporal mask. The resulting score can either be used for ranking and selection, self-refining or both. The VAE encoder is omitted for brevity.

2 Related Work

Physical plausibility in video generation. Recent video generative models [7, 5, 18, 34, 36, 39, 11, 31] have achieved impressive visual fidelity, yet they still frequently violate basic physical principles [17, 28]. Previous work explored three lines of work to solve problem. The first improves physical fidelity during training by introducing physics-aware data, supervision, or learning objectives [35, 8]. The second incorporates explicit physical structure or simulation into the generation process [23]. The final approach addresses the problem at inference time through external guidance, for example using VLM-based planning or reward signals from pretrained world models. VLIPP and GPT4motion use a VLM as physics reward [38, 25], while WMReward uses VJEPA-2 as an external latent world model to provide rewards for search and guided sampling [42].

Inference-time alignment for image generation. Inference-time alignment for diffusion models can be categorized into search-based optimization-based methods. Search-based approaches improve alignment by exploring multiple denoising trajectories and selecting or resampling candidates according to a reward, including Best-of-N selection, SMC-style resampling, and value-based decoding [21, 20, 26, 45, 15]. Optimization-based approaches directly modify the sampling trajectory or noise variables to increase the target reward, for example through reward guidance or direct noise optimization [33, 10, 40, 1, 9]. Collectively, these works show that substantial gains can be obtained at test time without retraining the generator. However, they typically rely on externally defined reward models or prompt-alignment objectives.

Latent reward modeling. Recent work has explored training video generation models as latent reward models, motivated by the limitations of VLM-based evaluators. In particular, learning rewards models by directly training on noisy latent states has been shown to better match latent-space generation and avoid the overhead of pixel-space VLM rewards [22, 27, 29]. Alternatively, LikePhys [41] curates a synthetic benchmark and evaluates models using preference Plausibility Preference Error (PPE) metric on curated valid-invalid video pairs. Unlike these approaches, which either train dedicated latent reward models or depend on curated simulated pairs, our method directly reuses the frozen generator's own denoising residual as a self-score for physical plausibility.

3 Method

We introduce **Proprio**, a training-free inference-time framework with two complementary stages: self-scoring and self-refinement (see Fig. 2). Self-scoring asks whether a generated latent video is consistent with the frozen generator's learned dynamics: samples that the model can explain under controlled perturbations should induce smaller and more stable flow residuals. Since large

video generators are trained on natural video dynamics, this learned-flow consistency can serve as an intrinsic proxy for physical plausibility. Self-refinement then uses this score as an objective to improve a sample by updating only a structured parameterization of its initial sampling noise.

The self-scoring signal is built from four components. (1) We perturb each generated latent at multiple scheduler timesteps and noise realizations, probing the sample across different noise regimes. (2) We measure the model's flow prediction residual, reusing the generator's own training objective as a self-consistency signal. (3) We optionally apply a dynamic spatiotemporal mask to focus the score on motion-relevant regions. (4) We aggregate residuals across timesteps with inverse-variance weighting, giving more weight to reliable estimates and downweighting perturbation-sensitive ones. The resulting score supports three inference-time modes: best-of-N search, self-refinement, and hybrid search-and-refine. We briefly review the flow-matching formulation in Appendix F.

3.1 Flow Prediction under Scheduler Perturbations

Let c denote the conditioning input (e.g., text prompt or input image). Let $z_0 \in \mathbb{R}^{T\times H\times W\times C}$ denote the latent representation of a generated video x_0. We consider a pretrained video generator with parameters θ, which predicts a velocity (flow) field

$$\hat{v}_\theta : \mathbb{R}^{T\times H\times W\times C} \times \mathcal{T} \times \mathcal{C} \to \mathbb{R}^{T\times H\times W\times C}, \tag{1}$$

where $\mathcal{T}$ denotes the set of scheduler timesteps.

Modern video generators define perturbations through a scheduler rather than a fixed analytical path. Given a timestep $t \in \mathcal{T}$ and noise $\epsilon \sim \mathcal{N}(0, I)$, we construct a perturbed latent

$$z_t = \mathcal{P}(z_0, \epsilon, t), \tag{2}$$

where $\mathcal{P}$ denotes the forward perturbation process.

The model predicts a flow field $\hat{v}_\theta(z_t, t, c)$, which is trained to match a target transport of the form

$$v^\star(z_0, \epsilon) \approx \epsilon - z_0. \tag{3}$$

This corresponds to the standard velocity parameterization and enforces local consistency of the learned transport under perturbations of real data.

3.2 Intrinsic Self-Scoring via Flow Residuals

Given a latent sample z_0, we probe its consistency by sampling ϵ and evaluating the model at perturbed states z_t. We define the flow residual

$$r(z_0, \epsilon, t; c) = \hat{v}_\theta(z_t, t, c) - v^\star(z_0, \epsilon), \tag{4}$$

and the per-timestep score

$$\ell_t(z_0, \epsilon; c) = \|r(z_0, \epsilon, t; c)\|_2^2. \tag{5}$$

This quantity corresponds to the flow-matching regression error evaluated at z_0, providing a local surrogate for alignment with the learned data distribution. Intuitively, it measures how consistently the learned transport explains the sample under perturbations: samples near well-modeled regions yield small residuals, while deviations from the learned distribution produce larger discrepancies.

Dynamic Spatiotemporal Masking. The score ℓ_t aggregates errors across the entire latent volume, including large static regions that are less informative for physical plausibility. To focus the signal on informative motion regions, we introduce a dynamic spatiotemporal mask $M \in [0, 1]^{T\times H\times W}$.

Let u index spatiotemporal locations. Refer to Appendix D for mask details. The masked score is

$$\ell_t^{\text{motion}}(z_0, \epsilon; c) = \frac{\sum_u M_u \, \|r_u(z_0, \epsilon, t; c)\|_2^2}{\sum_u M_u + \delta}, \tag{6}$$

where $\delta > 0$ ensures numerical stability.

This can be interpreted as importance weighting over spatiotemporal regions, emphasizing locations with higher expected relevance to motion and physical interactions, thereby improving sensitivity to physically meaningful deviations.

Multi-Perturbation, Multi-Timestep Estimation. A single evaluation of ℓ_t is sensitive to both timestep and noise. We therefore aggregate across multiple perturbations and timesteps.

Let $\mathcal{T} = \{t_1, \ldots, t_K\}$ and $\epsilon^{(n)} \sim \mathcal{N}(0, I)$ for $n = 1, \ldots, N$. For each timestep t_k, we compute

$$\mu_k = \frac{1}{N} \sum_{n=1}^{N} \ell_{t_k}(z_0, \epsilon^{(n)}; c), \tag{7}$$

$$v_k = \frac{1}{N} \sum_{n=1}^{N} \Big(\ell_{t_k}(z_0, \epsilon^{(n)}; c) - \mu_k \Big)^2 . \tag{8}$$

The mean μ_k estimates the residual strength at timestep t_k, while the variance v_k measures how stable that estimate is across perturbations. This separates the magnitude of the inconsistency from the reliability of the measurement.

Variance-Weighted Aggregation. Different timesteps can provide signals of different reliability: some noise levels expose meaningful inconsistencies, while others produce residuals that are highly sensitive to the sampled perturbation. We therefore weight timestep estimates by inverse variance:

$$w_k = \frac{1}{v_k + \varepsilon}, \quad \tilde{w}_k = \frac{w_k}{\sum_{j=1}^{K} w_j}, \tag{9}$$

where $\varepsilon > 0$ is a small constant for numerical stability.

We use v_k to measure the stability of the score at timestep t_k: larger v_k indicates that the score is more sensitive to the perturbation noise. Inverse-variance weighting therefore assigns higher weight to more stable timesteps and lower weight to noisier ones.

The final score is

$$S(z_0; c) = \sum_{k=1}^{K} \tilde{w}_k \, \mu_k. \tag{10}$$

For the motion-aware variant, we define S_{motion} analogously using ℓ_t^{motion}, and optionally combine both:

$$S_{\text{hybrid}}(z_0; c) = \lambda S_{\text{global}}(z_0; c) + (1 - \lambda) S_{\text{motion}}(z_0; c), \tag{11}$$

where $\lambda \in [0, 1]$. The global score preserves broad sample consistency, while the motion-aware score emphasizes dynamic regions; the hybrid score balances these two views.

3.3 Inference-Time Selection and Refinement

The Proprio score defines a generator-native objective over samples. It can therefore be used not only for selection, but also for refinement: rather than simply choosing the lowest-scoring candidate, we can directly optimize the sampling noise to reduce the score. We thus propose an inference-time self-refinement procedure that updates the initial noise while keeping the generator fixed.

The Proprio score defines a generator-native objective over samples. Given candidates $\{z_0^{(i)}\}$, we select $\hat{i} = \arg\min_i S(z_0^{(i)}; c)$. This provides a simple test-time scaling mechanism: generate multiple candidates and keep the one that is most internally consistent according to the frozen generator.

Beyond selection, the same score can be used for self-refinement. Instead of modifying the generator weights, we optimize a structured parameterization of the initial sampling noise. Let $\zeta_0 \sim \mathcal{N}(0, I)$ denote the base noise. We introduce per-channel parameters $\mu \in \mathbb{R}^C$ and $\sigma \in \mathbb{R}^C_+$, and define

$$\zeta = \mu + \sigma \odot \zeta_0, \tag{12}$$

where $\mu \in \mathbb{R}^C$ and $\sigma \in \mathbb{R}^C_+$. We denote by $z_0(\zeta; c)$ the latent generated from noise ζ under conditioning c.

We optimize

$$\mathcal{L}(\mu, \sigma) = S(z_0(\zeta; c); c) + \beta \mathcal{L}_{\text{KL}}, \tag{13}$$

where $\mathcal{L}_{\text{KL}} = \frac{1}{C} \sum_{d=1}^{C} \frac{1}{2} \left(\mu_d^2 + \sigma_d^2 - 1 - 2 \log \sigma_d \right)$.

Table 1: Inference-time search on VideoPhy2. We report Semantic Adherence (SA), Physical Commonsense (PC), and Joint accuracy on the standard and hard splits. Higher is better.

	Wan2.2 5B						**Turbo Wan14B**						**Hunyuan Video**					
	Standard			Hard			Standard			Hard			Standard			Hard		
Method	SA	PC	Joint	SA	PC	Joint	SA	PC	Joint	SA	PC	Joint	SA	PC	Joint	SA	PC	Joint
Random	65.4	58.1	48.6	31.7	44.4	19.4	67.2	68.3	56.1	32.2	47.8	23.3	61.1	75.0	50.0	29.4	61.1	20.0
Qwen 7B [3]	67.0	63.7	52.0	35.0	42.2	18.9	65.6	69.4	55.0	32.8	46.1	22.2	61.7	78.3	**52.8**	31.7	64.4	23.9
WMReward [42]	62.2	66.1	51.1	33.9	47.8	20.6	63.9	73.3	53.9	30.0	53.9	20.0	57.2	78.3	48.3	27.2	70.0	21.7
Ours - S_{global}	67.6	64.8	52.5	30.0	50.0	18.9	65.9	73.7	**56.4**	33.3	53.3	22.2	56.7	77.2	46.1	25.0	70.8	18.8
Ours - S_{motion}	66.5	68.2	**55.3**	34.4	51.7	**23.3**	64.8	74.3	55.9	33.9	57.8	**23.9**	58.3	**79.4**	50.6	22.9	70.8	18.8
Ours - S_{hybrid}	66.5	**69.8**	**55.3**	33.3	**52.8**	22.8	64.2	**74.9**	**56.4**	33.3	**58.3**	**23.9**	58.9	78.3	50.6	28.9	**71.7**	**25.0**

The Gaussian-affine parameterization gives refinement enough flexibility to improve the sample while keeping the optimized noise close to the generator's original Gaussian prior. Starting from $\mu = \mathbf{0}$ and $\sigma = \mathbf{1}$, we optimize $(\mu, \log \sigma)$ with Adam for a small number of steps while keeping the generator fixed. In practice, we return the sample achieving the lowest unregularized Proprio score during optimization. The full algorithm is provided in Appendix C.

4 Experiments

We evaluate Proprio in three inference-time settings: *search*, *refinement*, and *search-and-refine* (Sec. 3). Experiments are conducted on both text-to-video (T2V) and image-to-video (I2V) generation, using VideoPhy2 [4] and Physics-IQ [28], respectively, two benchmarks designed to assess physical plausibility and reasoning in generated videos. We include additional experiments in Appendix A.

Benchmarks For T2V, we use VideoPhy2 [4], which reports three metrics: Semantic Adherence (SA), Physical Commonsense (PC), and Joint (SA=1 and PC=1). We follow the default evaluation protocol where scores greater than or equal to 4 are mapped to 1 and all others to 0. Following previous works [16], we randomly sample 360 prompts equally across standard and hard set. For I2V, we use Physics-IQ [28], which directly reports a physics score.

Baselines. For *search*, we compare against (i) random selection, (ii) a VLM-based scorer using Qwen-7B [3], and (iii) WMReward [42], which ranks samples using VJEPA2 surprise scores. We follow their official implementation, scoring full video frames and selecting the lowest surprise-score sample using Vjepa2-G. For *refinement*, we compare against WMReward's guidance-based optimization method using the recommended guidance of 0.005 and update frequency of 5.

Implementation details. Unless otherwise stated, we generate 16 candidate samples per prompt. For fair comparison, all methods score the same candidate set. Experiments are run on 2 H100 GPUs. We use each model's default frame rate: 24 FPS for Wan2.2 and Hunyuan Video 1.5, and 16 FPS for Turbo Wan2.2. Refer to Appendix E for more details.

4.1 Improving Generation via Inference-Time Search

We evaluate inference-time scaling by generating 16 candidate samples and selecting the best one using Proprio. We conduct experiments across three video models, Wan2.2 5B [34], TurboDiffusion Wan2.2 14B [43], and Hunyuan Video 1.5 [36], in both (T2V) and (I2V) settings.

Text-to-Video. As shown in Table 1, Proprio consistently outperforms random selection and the VLM-based scorer on VideoPhy2, while WMReward remains a stronger baseline than Qwen-7B. Across models, Qwen-7B provides limited benefit: although it occasionally improves semantic adherence (SA), these gains do not translate into stronger physical commonsense (PC) or joint performance, consistent with prior observations [42]. Compared to WMReward, Proprio achieves further improvements in most settings, particularly on PC and joint metrics. We observe a clear tradeoff between SA and PC: global scoring tends to better preserve SA, whereas motion-aware and hybrid scoring more strongly improve physical consistency. Among these, the hybrid variant is the most robust, consistently achieving the best balance between SA and PC and the strongest joint performance. While improving physical plausibility can reduce SA, Proprio exhibits a milder tradeoff than prior approaches, with smaller SA degradation relative to the gains in PC and joint

scores. Overall, these results indicate that generator-native self-scoring provides a more effective signal for selecting physically plausible T2V samples than generic VLM-based evaluation, and is competitive with or stronger than WMReward.

Image-to-Video. As shown in Table 2, the I2V setting exhibits a different trend. The motion and hybrid variants of Proprio perform similarly, with motion-based scoring slightly stronger sometimes. This is consistent with the fact that I2V candidates are anchored to the same input image and are therefore already semantically aligned, making motion dynamics the primary factors distinguishing samples. While WMReward performs strongly across models, Proprio slightly improves over it on Wan2.2 5B and Turbo and achieves the best overall result on Hunyuan Video with the motion-aware score. Overall, these results show that generator-native self-scoring provides a strong and competitive criterion for selecting physically plausible I2V generations.

Table 2: Inference-time search results on Physics-IQ across three models. Higher is better.

Method	Wan2.2 5B	Turbo	Hunyuan Video
Random	29.45	31.25	31.63
Qwen 7B [3]	28.51	31.32	32.23
WMReward [42]	32.40	34.67	34.75
Ours - S_{global}	31.47	34.48	33.73
Ours - S_{motion}	32.25	34.65	**36.09**
Ours - S_{hybrid}	**32.46**	**35.13**	35.13

4.2 Refining Generations with Inference-Time Optimization

We evaluate inference-time refinement using Proprio as the optimization signal. To make gradient-based optimization tractable, we use a distilled model (TurboDiffusion Wan 14B) and backpropagate through a reduced number of denoising steps (4 instead of 50), significantly lowering computational cost while preserving a useful refinement signal. We conduct experiments in both text-to-video (T2V) and image-to-video (I2V) settings on the VideoPhy2 [4] and Physics-IQ [28] benchmarks.

As shown in Table 3, Proprio consistently improves over the base model and outperforms WMReward, with the largest gains observed when combining search and refinement. Refinement alone already yields meaningful improvements, and initializing from better candidates (via search) and then refining them leads to consistently stronger results, indicating that Proprio benefits from both selection and optimization. On Physics-IQ, the motion variant is the strongest, meaning that motion-focused refinement is very effective for I2V physical plausibility. On VideoPhy2, the hybrid score typically yields the strongest physical consistency on the standard split, while both motion and hybrid variants substantially improve performance on the hard split. Across settings, the most consistent gains are observed on physical commonsense (PC), indicating that refinement primarily improves physical realism. Overall, these results demonstrate that Proprio provides an effective optimization signal for challenging physical reasoning tasks, with motion-aware scoring particularly beneficial for I2V and hybrid scoring offering the most robust improvements for T2V.

4.3 Physical Plausibility Beyond Visual Quality

Proprio measures consistency with the generator's learned dynamics. Therefore, we investigate whether its self-score captures physical plausibility beyond generic visual quality, through human evaluation, FID-matched comparisons, and physically valid–invalid video pairs.

Human evaluation. To complement benchmark-based evaluation, we conduct a pairwise human evaluation study to test whether Proprio's self-score agrees with human judgments. We sample 64 video pairs for search and refinement settings across T2V and I2V, and collect 330 responses. In the search setting, each pair compares a low-score and high-score Proprio sample for the same condition. In refinement, each pair compares a Proprio-refined video with its unrefined baseline. As shown in Fig. 3, human preferences align well with Proprio on physical plausibility. Raters prefer the lower-score sample in 67.3% of search comparisons and the Proprio-refined sample in 66.7% of refinement comparisons. The effect is weaker for visual quality, where Proprio-selected or refined videos are preferred in roughly half of comparisons.

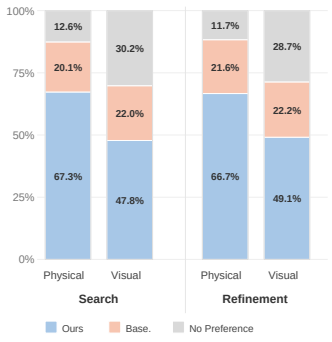


Figure 3: Human preference assessing visual quality and physical plausibility.

Table 3: Inference-time refinement on Turbo Wan2.2. We report Physics-IQ and the VideoPhy2 breakdown on the standard and hard subsets. Higher is better.

Method	Physics-IQ	VideoPhy2					
		Standard			Hard		
	Score	SA	PC	Joint	SA	PC	Joint
Turbo Wan2.2	32.2	65.6	67.2	55.0	31.1	45.6	19.4
Refinement only							
+ WMReward ($\mathcal{G}$)* [42]	32.4	67.3	67.2	55.6	32.8	45.0	21.7
+ Proprio ($\nabla\ S_{\text{global}}$)	34.6	65.4	68.2	54.2	29.4	43.9	20.0
+ Proprio ($\nabla\ S_{\text{motion}}$)	**35.8**	65.4	69.3	54.2	32.2	47.2	20.6
+ Proprio ($\nabla\ S_{\text{hybrid}}$)	34.9	65.6	**71.7**	54.4	32.2	**50.6**	21.1
Search (BoN4) + refinement							
+ WMReward ($\mathcal{G}$ + BoN4)* [42]	34.7	64.4	67.2	51.1	32.8	48.9	19.4
+ Proprio ($\nabla\ S_{\text{global}}$ + BoN4)	36.7	64.4	71.7	54.4	29.4	54.4	21.1
+ Proprio ($\nabla\ S_{\text{motion}}$ + BoN4)	**37.4**	66.1	71.7	52.8	29.7	54.3	21.1
+ Proprio ($\nabla\ S_{\text{hybrid}}$ + BoN4)	37.1	63.9	**73.3**	51.1	31.1	**55.0**	21.1
Search (BoN8) + refinement							
+ WMReward ($\mathcal{G}$ + BoN8)* [42]	35.5	66.7	68.9	52.8	30.6	52.2	21.1
+ Proprio ($\nabla\ S_{\text{motion}}$ + BoN8)	**37.5**	65.0	73.3	55.6	31.7	**54.4**	23.3
+ Proprio ($\nabla\ S_{\text{hybrid}}$ + BoN8)	37.3	66.1	**74.4**	57.8	29.6	54.2	20.7

* self reproduced. $\mathcal{G}$ indicates guidance

These results suggest that Proprio improves perceived physical plausibility while largely preserving visual quality.

Physical plausibility under matched FID. To further disentangle physical plausibility from generic visual quality, we conduct an additional evaluation over the entire Physics-IQs set. For each scenario, we manually compare higher and lower-ranked Proprio samples, constructing two groups with matched aggregate FID of 4.66. Despite having comparable FID, the higher-ranked Proprio videos are preferred in 68.7% of all comparisons and 79.5% of decisive comparisons after excluding "no preference.". These results indicate that Proprio's ranking captures physical plausibility cues that are not solely explained by differences in visual quality.

Discriminating physically valid and invalid videos. We additionally evaluate Proprio with Hunyuan Video on the LikePhys dataset [41], which contains synthetic physically valid–invalid video pairs designed to control for visual appearance. Proprio correctly assigns a lower self-score to the physically valid video in 71.0% of pairs, providing further evidence that its internal consistency signal is informative about physical validity.

These evaluations suggest that the generator's flow residual provides a useful empirical proxy for physical plausibility, rather than merely reflecting generic visual quality. We also provide additional human preference results and visual quality evaluations using VBench [14] in Appendix A.

4.4 Comparison with Existing Inference-Time Methods

We compare Proprio against existing inference-time methods, including WMReward [42], FreeInit [37], and Predict-and-Perturb (P&P) [16], on TurboWan2.2. For Proprio, N denotes the number of independent noise perturbations used to estimate the self-score at each timestep (Sec. 3). Increasing N provides a more robust score estimate at additional computational cost. Table 4 reports physical plausibility and wall-clock inference time for both I2V and T2V generation. With a single perturbation ($N = 1$), Proprio achieves higher Physics-IQ and Physical Commonsense (PC) than all evaluated methods. Increasing N to 4 further improves Physics-IQ from 34.60 to 35.80, although PC decreases. These results demonstrate a good performance–cost trade-off for generator-native

Table 4: Comparison with inference-time methods on TurboWan2.2. We report inference time per sample (seconds).

Method	I2V		T2V	
	Time	PhysIQ	Time	PC
TurboWan2.2	**37.54**	32.40	**14.36**	67.20
WMReward [42]	87.35	32.40	30.38	67.20
FreeInit [37]	129.22	32.42	101.86	68.89
P&P [16]	120.74	30.50	86.04	64.44
Proprio ($N = 1$)	79.67	<u>34.60</u>	28.93	**75.00**
Proprio ($N = 4$)	166.38	**35.80**	57.93	<u>71.70</u>

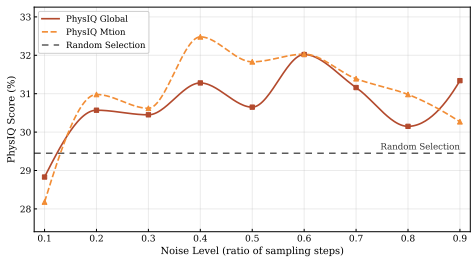


Figure 4: Effect of noise levels at different timestep ranges on Physics-IQ for Wan2.2 5b.

Range	Method	VideoPhy2 SA	PC	Joint
All noise (0.2–0.8)	S_{global}	71.66	60.56	51.66
	S_{motion}	69.44	61.66	51.11
Mid noise (0.3–0.6)	S_{global}	67.04	64.25	51.40
	S_{motion}	67.03	66.48	56.42
	S_{global} + var	67.59	64.80	52.51
	S_{motion} + var	66.48	**68.16**	55.31

Table 6: Effect of aggregating scores at different noise levels and with variance weighting on scoring for VideoPhy2 (standard) with Wan2.2 5B.

self-refinement. Additional perturbation-budget experiments and inference-cost comparisons are provided in Appendix A.3.

5 Ablation Studies

Effect of noise levels for scoring. As shown in Fig. 4 and Table 6, the most informative scoring signal arises from an intermediate noise regime, rather than from very low or very high noise levels. On Physics-IQ, both global S_{global} and motion aware S_{motion} scoring improve substantially when evaluation is restricted to mid-noise timesteps, while performance degrades at the extremes. A similar pattern holds on VideoPhy2: aggregating scores over the full noise range yields the highest semantic adherence (SA), but significantly weaker physical commonsense (PC) and joint performance. In contrast, restricting evaluation to mid-noise timesteps improves PC and produces the strongest joint scores. These results indicate that extreme noise levels yield less reliable or less informative signals, whereas mid-noise perturbations provide the most discriminative signal for physical plausibility.

Effect of variance weighting. Table 6 shows that inverse-variance weighting improves score aggregation by emphasizing timesteps with more stable residual estimates. This effect is most pronounced for the motion score, where variance weighting increases PC from 66.48 to 68.16, yielding the best overall PC result. Gains for the global score are smaller but consistent, indicating that not all timesteps contribute equally reliable information. These results highlight that effective self-scoring depends not only on selecting an appropriate noise regime, but also on weighting timesteps according to their stability. In practice, combining mid-noise with variance-weighted motion scoring provides the strongest and most robust performance. We note that this effect is significantly weaker on distilled models, where variance weighting yields only marginal improvements (see Appendix A.4).

Dynamic masking strategy. We explore different setups for the motion-based masking. Table 5 compares three dynamic masking variants for optimization. The *motion* mask is the default introduced in Sec. 3. The *textaware* mask derives motion from cross-attention, making it explicitly prompt-grounded, while the *attention-based* mask uses self-attention feature changes as an internal proxy for motion, similar to [32]. Additional details are provided in Appendix D. All three variants provide useful dynamic signals, but exhibit distinct tradeoffs. The motion mask achieves the strongest overall performance, with the best Physics-IQ score and the highest PC and joint scores on VideoPhy2, indicating that direct motion cues are the most reliable signal in this setting. The textaware mask yields the highest semantic adherence (SA), suggesting that prompt-grounded masking better preserves semantic alignment, but at some cost to physical consistency. The attention-based variant remains competitive but is slightly weaker. Overall, these results show that Proprio is compatible with multiple forms of dynamic masking, and that the choice of mask controls the tradeoff between semantic adherence and physical consistency.

Table 5: Dynamic masking strategies for optimization on Turbo Wan2.2.

Masking	Physics-IQ Score	VideoPhy2 Standard SA	PC	Joint
Motion	**35.8**	65.36	**69.27**	**54.19**
Attention-based	34.10	63.33	68.33	52.22
Textaware	33.04	**66.66**	65.56	53.33

6 Discussion

Importance of mid-level noise. Our ablations in Section 5 show that the most informative self-scoring signal arises at intermediate noise levels. At very low noise, the perturbed latent remains too close to the original sample, making the residual largely local and less sensitive to higher-level violations of physical plausibility. At very high noise, much of the sample-specific structure is lost, so predictions are driven more by the model's generic prior and become less discriminative. Mid-level noise provides the best balance: it preserves enough structure for the score to remain sample-specific while perturbing the latent enough to reveal inconsistencies in spatiotemporal dynamics. This explains the stronger PC and joint performance of mid-noise evaluation relative to low or high-noise regimes.

Inference scaling versus inference refinement. Comparing search and refinement reveals a clear dependence on the generation setting. In T2V, search is particularly effective because samples from the same prompt exhibit substantial scene and motion diversity, making selection from a larger candidate pool highly beneficial. This aligns with prior works [12] discussing the diversity-performance tradeoff in T2V. In contrast, I2V candidates are anchored to the same input image, reducing semantic diversity and leaving motion quality as the primary source of variation. As a result, the advantage of large-scale search diminishes, and the gap between search and refinement narrows. In this regime, motion-aware refinement becomes effective even with a smaller search budget. Overall, these results indicate that scaling search is more important for T2V, whereas in I2V the marginal benefit of additional diversity is lower and refinement plays a stronger role.

7 Conclusion and Limitations

We introduced **Proprio**, a generator-native inference-time framework for improving physical plausibility in video generation. By reusing the frozen generator's own denoising dynamics as an intrinsic self-scoring signal, Proprio enables both search and refinement without external reward models. Across text-to-video and image-to-video benchmarks, Proprio consistently improves physical plausibility and provides a strong alternative to external evaluators. Proprio is inherently limited by the knowledge encoded in the base generator, it cannot recover dynamics that the model has not learned. In addition, like other inference-time scaling methods, Proprio increases inference cost by requiring additional candidate scoring, search, or refinement steps. Our refinement experiments mainly rely on distilled models for computational tractability, and scaling gradient-based refinement to larger, non-distilled generators remains challenging due to the cost of backpropagation through long denoising trajectories. An important direction for future work is therefore the development of more efficient scoring and refinement. Overall, our results suggest a shift in perspective for video generation: rather than relying solely on external supervision or auxiliary reward models, future systems may increasingly exploit generator-internal signals for self-evaluation and self-alignment.

Acknowledgments and Disclosure of Funding

This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a144. Kaouther Messaoud is supported by Hi! PARIS and State funding managed by the French National Research Agency (ANR) under the France 2030 program, reference ANR-23-IACL-0005.

References

[1] 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] 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. *arXiv preprint arXiv:2506.09985*, 2025.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] 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. In *International Conference on Learning Representations*, 2026.

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

[6] 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.

[7] 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. *OpenAI Blog*, 2024. Online; accessed from OpenAI.

[8] Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, Xiaoliang Dai, Xuan Ju, Alan Yuille, and Ji Hou. Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. In *European Conference on Computer Vision*, pages 91–109. Springer, 2026.

[9] 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. In *Advances in Neural Information Processing Systems*, 2025.

[10] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9380–9389, 2024.

[11] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[12] Mariam Hassan, Bastien Van Delft, Wuyang Li, and Alexandre Alahi. Anchored video generation: Decoupling scene construction and temporal synthesis in text-to-video diffusion models. *arXiv preprint arXiv:2512.16371*, 2026.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2024.

[15] Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

[16] Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-refining video sampling. In *Forty-third International Conference on Machine Learning*, 2026.

[17] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning*, pages 28991–29017, 2025.

[18] 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. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[19] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, 2023.

[20] 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.

[21] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *Advances in Neural Information Processing Systems*, 2025.

[22] Gongye Liu, Bo Yang, Yida Zhi, Zhizhou Zhong, Lei Ke, Didan Deng, Han Gao, Yongxiang Huang, Kaihao Zhang, Hongbo Fu, and Wenhan Luo. Beyond VLM-based rewards: Diffusion-native latent reward modeling. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 79354–79376, 2026.

[23] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024.

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[25] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1430–1440. IEEE, 2024.

[26] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2523–2534, 2025.

[27] Xiaoyue Mi, Wenqing Yu, Jiesong Lian, Shibo Jie, Ruizhe Zhong, Zijun Liu, Guozhen Zhang, Zixiang Zhou, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Fan Tang. Video generation models are good latent reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 4719–4728, 2026.

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

[29] Felipe Nuti, Tim Franzmeyer, and João F. Henriques. Extracting reward functions from diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

[30] NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

[31] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter

Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[32] Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22911–22921. IEEE, 2025.

[33] Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. In *Forty-second International Conference on Machine Learning*, 2025.

[34] 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.

[35] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. In *Advances in Neural Information Processing Systems*, 2025.

[36] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[37] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European conference on computer vision*, pages 378–394. Springer, 2024.

[38] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12360–12370. IEEE, 2025.

[39] 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*, pages 83048–83077, 2025.

[40] 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. In *Advances in Neural Information Processing Systems*, pages 22370–22417, 2024.

[41] 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*, 2026.

[42] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16118–16129, 2026.

[43] Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025.

[44] 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.

[45] XiangCheng Zhang, Haowei Lin, Haotian Ye, James Y Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. In *International Conference on Learning Representations*, pages 39133–39166, 2026.

A Additional experiments and ablations

A.1 Diagnostic Study: Preference for natural video dynamics

To test whether the self-score captures a useful preference for natural video dynamics, we conduct a small but highly curated diagnostic check on Physics-IQ. We select 45 *distinct* scenarios from Physics-IQ and use their ground-truth videos as physically plausible references. For each case, we construct a matched physical-failure example using Turbo Wan2.2, either by prompting the model toward a specific implausible behavior or by generating multiple samples and selecting one with a clear physics violation. We manually curate these pairs so that the generated counterpart remains visually and semantically close to the reference where possible, while the main intended difference is physical plausibility. The physical failures include, unrealistic collisons, rigid objects changing shapes, unrealistic shadows, liquid dynamics, unrealistic falls, wrong reaction to magnet and more. We then score both the ground-truth video and the generated failure using the same Turbo Wan2.2 model with the motion Proprio loss, and compare which video is preferred by the self-score.

The ground-truth video is preferred in 82.2% of pairs (37/45), suggesting that the generator-native residual tends to favor natural video dynamics over the model's own physically implausible generations. This study is intended only as a diagnostic sanity check supporting our interpretation that learned-distribution consistency can serve as a useful proxy for physical plausibility. We show examples of the curated pairs in Figures 11 and 12.

A.2 Visual Quality

To assess whether the gains in physical plausibility come at the expense of visual quality, we additionally evaluate the generated videos on VBench [14]. As shown in Table 7, Proprio does not degrade perceptual quality across either T2V or I2V. Subject consistency, background consistency, motion smoothness, aesthetic quality, and imaging quality remain close to the baseline across refinement, search, and search+refinement settings. In some cases, these metrics slightly improve, suggesting that Proprio's physical-plausibility gains are not obtained by sacrificing visual quality. Overall, these results indicate that the improvements in physical plausibility obtained by Proprio do not come at the cost of perceptual video quality, and may in some cases slightly improve it.

Table 7: VBench quality metrics for Turbo Wan2.2 under baseline sampling, Proprio refinement (∇), search (BoN), and search + refinement (∇ + BoN).

Method	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality
		VideoPhy2 (T2V)			
Turbo Wan2.2	0.92	0.92	0.99	0.51	0.67
+ Proprio(∇)	0.93	0.93	0.98	0.51	0.68
+ Proprio(BoN)	0.95	0.95	0.99	0.54	0.68
+ Proprio(∇ + BoN)	0.93	0.93	0.98	0.52	0.68
		PhysicsIQ (I2V)			
Turbo Wan2.2	0.95	0.96	0.99	0.49	0.66
+ Proprio(∇)	0.95	0.96	0.99	0.49	0.66
+ Proprio(BoN)	0.95	0.96	0.99	0.48	0.68
+ Proprio(∇ + BoN)	0.96	0.96	0.99	0.49	0.66

A.3 Inference-Time Cost and Perturbation Budget

Effect of the perturbation budget. Proprio estimates its self-score by averaging residuals over N independent noise perturbations at each selected timestep. We investigate how the perturbation budget affects refinement performance on TurboWan2.2. As shown in Table 8, increasing N from 1 to 4 improves Physics-IQ from 34.60 to 35.80. However, Physical Commonsense (PC) and Joint accuracy do not improve monotonically as $N = 2$ achieves the highest PC, while $N = 1$ achieves the highest Joint accuracy. One possible explanation lies in the different conditioning settings. In I2V, the input image constrains the generated scene, potentially making differences in physical plausibility across samples more subtle and requiring additional perturbations to obtain a reliable self-score. In

contrast, T2V allows greater variation in appearance and motion, where fewer perturbations may suffice to distinguish plausible dynamics. These results suggest that the optimal perturbation budget may depend on the task and generation setting.

Table 8: Effect of the number of noise perturbations N on Proprio refinement using TurboWan2.2. Physics-IQ is evaluated on I2V, while SA, PC, and Joint are evaluated on T2V. Higher is better.

Perturbations	PhysIQ	SA	PC	Joint
$N = 1$	34.60	66.66	75.00	**56.67**
$N = 2$	34.80	65.00	**75.56**	53.33
$N = 3$	34.90	64.44	75.00	49.44
$N = 4$	**35.80**	65.60	71.70	54.40

Effect of the Search Candidate Budget Increasing the candidate budget from $B = 4$ to $B = 16$ consistently improves Physics-IQ and PC (Table 9). Search cost grows almost linearly with the number of generated and scored candidates.

Table 9: Effect of the search candidate budget B. Higher is better.

Selection	Physics-IQ	SA	PC	Joint
Random selection	31.25	67.20	68.30	56.10
Proprio, $B = 4$	33.87	65.90	72.22	56.98
Proprio, $B = 8$	34.74	64.80	73.18	55.86
Proprio, $B = 16$	**35.13**	64.20	**74.90**	56.40

Inference-time scaling and computational cost. We further analyze the cost of Proprio's scoring procedure. For best-of-B search, generating one candidate takes 37.54 s for I2V and 14.36 s for T2V. With the default perturbation budget $N = 4$, Proprio adds 24.96 s of forward-only scoring per candidate, resulting in total costs of 62.50 s and 39.32 s per candidate for I2V and T2V, respectively. Consequently, total search time scales approximately linearly with the candidate budget B.

A.4 Effect of variance on distilled model

Table 10 shows that variance weighting has only a limited effect on the distilled Turbo model. On Physics-IQ, using variance weighting does not improve performance and in fact leads to slightly lower scores for the global and combined variants, while leaving the dynamic variant unchanged. A similar pattern appears on VideoPhy2 where the differences are small across all three variants, and only minor shifts between semantic adherence and physical consistency. In particular, variance weighting slightly improves PC for the global and combined variants, but these gains are offset by lower SA or joint performance, and the dynamic variant is nearly unchanged. This contrasts with the non-distilled Wan model, where variance weighting provides clearer gains. One possible explanation is that distillation compresses the denoising trajectory into fewer steps, making the timestep-wise residual estimates more uniform in reliability. We therefore, do not use variance weighting for optimization setting as it's done on the distilled model.

Table 10: Effect of variance weighting on the distilled Turbo model.

	Physics-IQ		VideoPhy2 Standard					
Variant	w/o variance	w/ variance	SA	PC	Joint	SA	PC	Joint
			w/o variance			w/ variance		
S	34.5	34.3	66.7	73.3	58.9	65.9	73.7	56.4
S_{motion}	34.6	34.6	65.0	74.4	56.7	64.8	74.3	55.9
S_{hybrid}	35.1	34.6	64.4	74.4	56.1	64.2	74.9	56.4

A.5 Motion-stratified results

While motion is an important aspect of physical plausibility, its magnitude alone is not a proxy for physical correctness. A physically plausible video may contain little motion when the scene demands it, whereas an implausible generation may contain excessive or inconsistent motion. We therefore evaluate whether Proprio improves physical plausibility among samples with comparable motion magnitude. We compute a motion magnitude score per video and split videos into quartiles within each dataset for TurboWan2.2. The low-motion bin contains videos up to the 33rd percentile, the medium-motion bin contains videos between the 33rd and 67th percentiles, and the high-motion bin contains videos above the 67th percentile. This data-adaptive binning avoids arbitrary thresholds and enables fair within-dataset comparisons across motion regimes. As shown in Table 11, Proprio-selected samples generally improve benchmark performance within motion bins. On Physics-IQ, Proprio outperforms random selection, with particularly large gains in the low- and high-motion regimes. On VideoPhy2, Proprio improves the PC metric in the low- and medium-motion bins and remains comparable in the high-motion bin. These results indicate that Proprio's gains are not explained solely by changes in motion magnitude; instead, the score continues to identify more physically plausible generations among videos with similar motion levels.

Table 11: Motion-stratified benchmark control. Within each motion bin, we compare Proprio samples against random samples. For Physics-IQ, we report benchmark score; for VideoPhy2, we report PC metric. Positive Δ indicates Proprio-selected samples outperform random samples.

Benchmark	Motion bin	Proprio	Random	Δ
Physics-IQ	Low	**32.0**	24.8	+7.2
	Medium	**33.0**	**33.0**	+0.0
	High	**41.0**	33.6	+7.4
VideoPhy2	Low	**81.9**	70.8	+11.1
	Medium	**78.9**	77.4	+1.5
	High	58.0	**59.4**	-1.4

A.6 Robustness Across Random Seeds

We repeat Wan2.2 5B search and Turbo Wan2.2 refinement over three random seeds. Table 12 reports the mean and standard deviation. Proprio and WMReward achieve comparable Physics-IQ performance in search, while Proprio obtains higher PC. For refinement, Proprio improves Physics-IQ and PC, with differences exceeding the observed across-seed standard deviations, although SA and Joint decrease.

Table 12: Mean $\pm$ standard deviation across three seeds. Physics-IQ evaluates I2V; SA, PC, and Joint evaluate T2V. Higher is better.

Method	Physics-IQ	SA	PC	Joint
Wan2.2 5B: search				
WMReward	32.45 ± 0.05	61.67 ± 0.55	66.93 ± 0.84	50.82 ± 0.28
Proprio	32.50 ± 0.04	66.93 ± 0.45	69.20 ± 0.60	55.63 ± 0.35
Turbo Wan2.2: refinement				
WMReward	32.33 ± 0.11	67.26 ± 0.34	66.29 ± 1.60	54.46 ± 1.94
Proprio	35.59 ± 0.36	64.04 ± 1.99	71.33 ± 1.48	52.27 ± 1.86

A.7 Comparison with Plain Denoising Loss

We compare Proprio with optimizing the generator's plain denoising loss under the same Turbo Wan2.2 refinement setup. As shown in Table 13, Proprio improves all four metrics, suggesting that its gains are not explained by optimizing plain denoising loss alone.

Table 13: Refinement objectives on Turbo Wan2.2. Higher is better.

Objective	Physics-IQ	SA	PC	Joint
Plain denoising loss	33.04	64.44	67.78	51.67
Proprio	**35.80**	65.60	**71.70**	54.40

A.8 Additional Refinement Example on HunyuanVideo

As an additional refinement example, we apply Proprio to HunyuanVideo, backpropagating through 8 denoising steps and using inverse-variance weighting. Table 14 shows improvements in Physics-IQ and PC, accompanied by lower SA and Joint. This provides only an initial evidence that Proprio can guide refinement beyond the 4-step distilled models, while efficient refinement through longer sampling trajectories remains future work.

Table 14: Refinement on HunyuanVideo with 8 backpropagated denoising steps. Higher is better.

Method	Physics-IQ	SA	PC	Joint
Baseline	31.50	62.22	76.66	52.77
Proprio	**32.20**	60.22	**77.84**	51.70

A.9 Diagnostic of Inverse-Variance Weighting

To test whether inverse-variance weighting can hide failure signals, where a candidate receives a favorable score because high-residual timesteps are assigned small weights, we analyze 3,166 candidates from Wan2.2 5B Physics-IQ search. The weighted score remains strongly correlated with the uniform timestep mean (Spearman's $\rho = 0.974$), the 75th-percentile timestep residual ($\rho = 0.960$), and the worst-timestep residual ($\rho = 0.927$).

Among candidates selected by the weighted score, 93.4% rank in the better half at all four individual timesteps, and 99.0% rank in the better half at three or more timesteps. Moreover, 85.4% have a worst-timestep residual in the best quartile of their candidate set, and 94.4% remain in the better half even when ranked by their worst-timestep residual. These results indicate that favorable rankings are generally not produced by suppressing isolated high-error timesteps. Selected candidates typically have low raw residuals across the informative probes.

A.10 Human Evaluation

We conduct a human evaluation with overall results presented in Sec 4.3 but we provide more results in this section. Figure 5 provides the human evaluation results split by generation setting. The trend is consistent across both I2V and T2V: voters prefer Proprio for physical plausibility in the majority of comparisons, with 67.9% preference in I2V and 66.3% in T2V. This suggests that Proprio's improvements are perceived by humans in both conditional image-to-video and text-to-video settings. For visual quality, the pattern is more mode-dependent. In I2V, most annotators report no preference, which is expected because both videos are anchored by the same input image and therefore often share very similar scene content and appearance. In T2V, visual preferences are more decisive, with 55.3% favoring Proprio, likely because generated scenes can differ more substantially in terms of composition. Overall, the human study indicates that Proprio's main benefit is improved perceived physical plausibility, while visual quality is generally preserved and sometimes preferred.

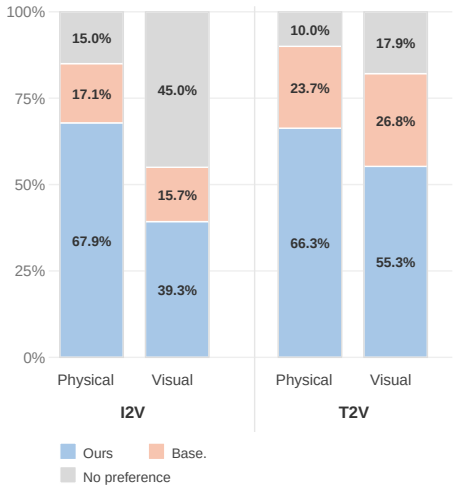


Figure 5: Human preference results split by generation setting.

We additionally provide a screenshot example of the study with full instructions given to participants in figure 6. Participants volunteered to answer the questions with no compensation. No personal data were collected, only answers to the questions of study.

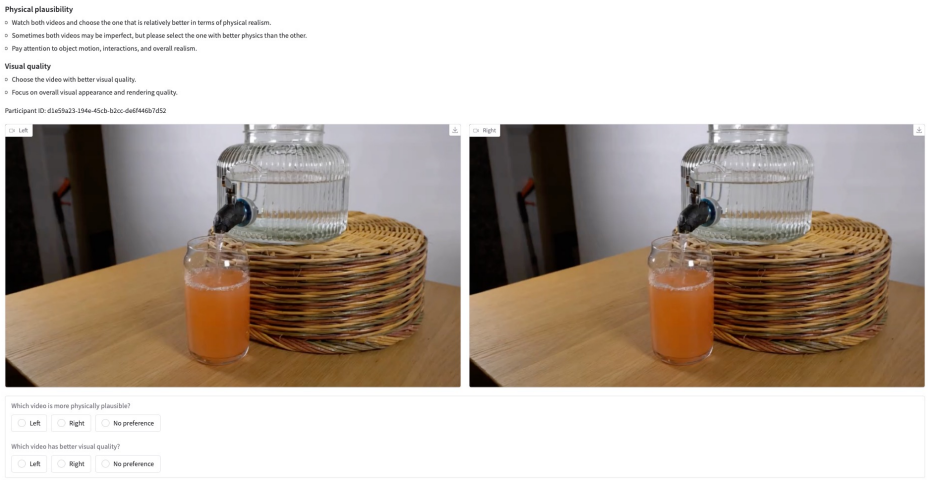


Figure 6: A screenshot from the human evaluation study showing full instructions and layout used.

B Additional Discussion

Semantic Adherence - Physical Commonsense tradeoff. Improving physical plausibility in T2V is inherently multi-objective: candidates that faithfully realize complex prompts may involve difficult interactions and therefore be more prone to physical errors, while physically safer generations may simplify the requested action or reduce interaction complexity. This has also been observed in prior works [42]. While Proprio exhibits similar tradeoff, some preserve semantic adherence better; for

example S_{global} generally preserves SA better while its improvement on physics commonsense is slightly less than S_{motion} as seen in Section 4. Similarly, the text-aware dynamic mask in Section 5 better preserves semantic adherence, at the cost of slightly weaker physical improvements. In practice, the choice of score can therefore be matched to the application: S_{global} is preferable when prompt fidelity is critical, S_{motion} when physical dynamics are the main concern, and S_{hybrid} as a balanced default for improving physical plausibility while maintaining semantic alignment.

C More Details on optimization

Algorithm 1 summarizes the self-refinement procedure used in Sec. 3. Starting from the original sampling seed, we optimize a Gaussian-affine reparameterization of the initial noise, $\zeta = \mu + \sigma \odot \zeta_0$, while keeping the video generator fixed. At each iteration, the current noise is used to generate a latent sample $z_0(\zeta; c)$, which is evaluated with the Proprio score together with a KL regularizer that keeps the refined noise close to the standard Gaussian prior. The optimization is performed over $(\mu, \log \sigma)$ using Adam, and the final output is the sample achieving the lowest unregularized Proprio score during refinement. This parameterization constrains the update to a structured family of perturbations, which stabilizes optimization and reduces drift away from the generator's original sampling distribution.

Algorithm 1 Proprio Self-Refinement Algorithm

Sample base noise $\zeta_0 \sim \mathcal{N}(0, I)$ from the initial seed.
Initialize $\mu \leftarrow \mathbf{0}$ and $\log \sigma \leftarrow \mathbf{0}$.
for $n = 1$ **to** N **do**
  $\sigma \leftarrow \exp(\log \sigma)$
  $\zeta \leftarrow \mu + \sigma \odot \zeta_0$
  Generate latent sample $z_0(\zeta; c)$ with the frozen generator
  Evaluate Proprio score $S(z_0(\zeta; c); c)$
  Compute

$$\mathcal{L}_{\text{KL}} = \frac{1}{C} \sum_{d=1}^{C} \frac{1}{2} \left(\mu_d^2 + \sigma_d^2 - 1 - 2 \log \sigma_d \right)$$

  Form the optimization objective

$$\mathcal{L}(\mu, \sigma) = S(z_0(\zeta; c); c) + \beta \mathcal{L}_{\text{KL}}$$

  Update $(\mu, \log \sigma)$ with Adam (optional gradient clipping)
end for
return the sample $z_0(\zeta^\star; c)$ achieving the lowest unregularized Proprio score $S(z_0(\zeta; c); c)$ during optimization

D Dynamic masking strategies

We consider three dynamic masking strategies in ablations 5, each designed to emphasize regions that are most informative for physical plausibility. All three produce a spatiotemporal saliency volume $S \in \mathbb{R}^{T \times H \times W}$ that is later converted into the latent dynamic mask used for scoring and refinement. The variants differ only in how this saliency volume is estimated.

Motion mask. The motion mask is derived from temporal changes in the generated samples. Given a sequence of spatiotemporal features or decoded frames $\{Y_t\}_{t=1}^{T}$, we compute framewise motion maps using a temporal difference operator,

$$D_t = \phi(Y_t, Y_{t-1}), \tag{14}$$

where ϕ measures local temporal change. The resulting maps are normalized and projected onto the latent grid to obtain a latent saliency volume S_{motion}. Intuitively, this mask highlights regions with significant temporal variation, thereby emphasizing object motion, interaction changes, and potential dynamic inconsistencies.

Textaware mask. The textaware mask derives motion from cross-attention, making it explicitly prompt-grounded. For selected transformer layers and probe timesteps, we extract the cross-attention affinity between latent query tokens and prompt tokens, average over the relevant text tokens, and reshape the result into a spatiotemporal affinity map

$$A_t \in \mathbb{R}^{H \times W \times T}. \tag{15}$$

We then convert this affinity into a motion-like signal by measuring its temporal variation,

$$S_{\text{text}}(t) = |A_t - A_{t-1}|, \tag{16}$$

which highlights regions whose prompt-conditioned attention changes over time. In this way, the mask focuses on dynamic regions that are both visually active and semantically relevant to the text prompt.

Attention-based mask. The attention-based mask uses internal self-attention feature changes as a proxy for motion, similar in spirit to prior work using attention dynamics as an internal motion signal [32]. For selected self-attention layers, we extract latent query and key features at consecutive timesteps and measure their temporal similarity at each spatial location. Let $q_t(h, w)$ and $k_{t-1}(h, w)$ denote the corresponding features. We define the saliency as

$$S_{\text{attn}}(t, h, w) = \text{ReLU}\big(1 - \cos\big(q_t(h, w), k_{t-1}(h, w)\big)\big)\,. \tag{17}$$

Locations whose internal features change more strongly over time therefore receive higher saliency. Unlike the other mask variants, this masking uses the model's own internal temporal feature evolution as a learned motion proxy.

Summary. The three masking strategies reflect different notions of dynamic saliency: the motion mask uses explicit visible motion, the textaware mask emphasizes prompt-grounded temporal changes, and the attention-based mask uses internal self-attention feature changes as a model-centric proxy for motion. As shown in Table 5, all three provide useful dynamic signals, but induce different tradeoffs between semantic adherence and physical consistency.

E More Implementation Details

Unless otherwise stated, Proprio scoring is computed in latent space over a fixed set of scheduler timesteps sampled from the mid-noise regime, using multiple perturbation samples per timestep and inverse-variance weighting across timesteps. For refinement, we backpropagate through a reduced number of denoising steps while keeping the generator frozen, and optimize only the Gaussian-affine parameters of the initial noise. Our default motion score uses the motion-based spatiotemporal mask described in Sec. 3, which is fixed during scoring and refinement unless otherwise specified. For hybrid search-and-refinement, all methods use the same candidate pool before refinement.

Table 15 summarizes the optimization hyperparameters used for Proprio self-refinement. Unless otherwise stated, we optimize the Gaussian-affine noise parameters for 10 steps using Adam with a learning rate of 2×10^{-3}. We apply gradient clipping for stability and use a small KL regularization term to keep the optimized noise close to the standard Gaussian prior. For score estimation during refinement, we average over 4 noise samples per timestep, which provides a more stable optimization signal while keeping the computational cost manageable. In all experiments, μ and σ are calculated per channel.

Table 15: Optimization hyperparameters used for Proprio self-refinement.

Hyperparameter	**Value**
Optimization steps (K)	10
Learning rate	2×10^{-3}
Gradient clipping	3×10^{-4}
KL regularization weight (β)	5×10^{-3}
Number of noise samples per timestep	4

For all experiments, we use 4 mid-noise levels linearly spaced between 0.3-0.6 of the sampling steps. For non-distilled models, variance weighting is used by using 4 random iterations per noise level and calculating variance in prediction. For the hybrid loss, $\lambda = 0.2$ is used for all models.

F Flow Matching Background

Flow matching formulates generation as learning a time-dependent vector field that transports a simple base distribution to the data distribution in latent space. Let z_1 denote a clean video latent and let $z_0 \sim \mathcal{N}(0, I)$ denote a Gaussian noise sample. A continuous path between noise and data is defined by

$$z_t = (1-t)z_0 + tz_1, \qquad t \in [0,1], \tag{18}$$

with corresponding target velocity

$$v^\star(z_0, z_1) = z_1 - z_0. \tag{19}$$

A conditional generator $\hat{v}_\theta(z_t, t, c)$ is then trained to predict this transport field by minimizing

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{z_1, z_0, t}\left[\|\hat{v}_\theta(z_t, t, c) - (z_1 - z_0)\|_2^2\right]. \tag{20}$$

In practice, modern video generators implement this perturbation through a scheduler rather than an explicit analytic path, so we write the noisy latent more generally as

$$z_t = \mathcal{P}(z_1, \epsilon, t), \qquad \epsilon \sim \mathcal{N}(0, I), \tag{21}$$

and the corresponding target transport under the standard velocity parameterization as

$$v^\star(z_1, \epsilon) \approx \epsilon - z_1. \tag{22}$$

At inference time, sampling starts from Gaussian noise and iteratively applies the learned vector field across timesteps to produce a latent video. In our method, we reuse this same flow-prediction objective at test time: by evaluating how well the frozen model predicts its target transport on perturbed generated samples, we obtain an intrinsic signal for ranking and refinement.

G Extended related work

Physical plausibility in video generation. Recent video generative models [7, 5, 18, 34, 36, 39, 11, 31, 28, 17, 28] have achieved impressive visual fidelity and temporal coherence, but multiple recent studies show that strong perceptual quality does not imply strong physical understanding or physically consistent dynamics. Diagnostic works such as Physics-IQ and related physical-law evaluations highlight persistent failures in collisions, object interactions, support relations, and generalization of physical behavior, suggesting that current video generators often reproduce surface realism without reliably capturing underlying physical structure [28, 17, 28]. Existing approaches to address this problem largely fall into three groups. A first line of work improves physical fidelity during training by introducing physics-aware data, supervision, or learning objectives, for example through physics-focused pretraining, preference optimization, or curated physics-aware corpora [35, 8]. A second line of work incorporates explicit physical structure or simulation into the generation process itself. For example, PhysGen uses rigid-body simulation to ground image-to-video generation, while GPT4Motion leverages GPT-4 to script Blender-based physical motions before video synthesis [23, 25]. A third line of work addresses physical plausibility at inference time through external guidance. VLIPP uses a vision-language model as a coarse motion planner with physics-aware reasoning, and WMReward uses VJEPA-2 as an external latent world model to score candidates and guide sampling toward physically plausible videos [38, 42]. Compared with these approaches, our method does not introduce external simulators, planners, or world-model rewards; instead, it reuses the frozen generator's own latent residual dynamics as a generator-native self-scoring signal for inference-time ranking and refinement.

Inference-time alignment for diffusion models. Inference-time alignment methods for diffusion models can be grouped into two broad families: *search-based* methods and *optimization-based* methods. Search-based methods improve alignment by spending additional compute on candidate exploration, trajectory branching, or resampling, followed by reward or value-based selection. This includes Best-of-N style inference-time scaling, value-based decoding, and dynamic search or resampling procedures that explicitly trade off exploration and exploitation during denoising [21, 20, 26]. Optimization-based methods instead directly modify the sampling trajectory or the underlying noise variables to improve a target objective. In particular, DNO optimizes injected noise online during the denoising process with respect to a reward [33], while InitNO optimizes the initial latent

using attention-derived prompt-alignment criteria and a distribution-preserving regularizer [10]. More generally, training-free guidance methods steer diffusion trajectories at test time using external objectives without retraining the generator [40, 1, 9]. Collectively, these works show that large gains can be achieved at inference time alone, without changing model weights. However, they typically depend on externally defined rewards, alignment heuristics, or verifier signals. In contrast, Proprio uses a generator-native objective: it scores and refines samples directly through the frozen video generator's own latent residual dynamics, avoiding external reward models.

Reward extraction from diffusion models. A broader line of work has explored whether diffusion models contain reward-like information that can be extracted or repurposed. In decision-making, Nuti et al. [29] define a *relative reward function* between a base and an expert diffusion model, and learn an auxiliary reward network by aligning its gradient to the difference in the two models' outputs. Their setting is primarily offline RL and control, but it provides an early indication that diffusion dynamics may encode preference-relevant structure beyond standard sample generation.

Latent reward modeling for generation. More recent work has studied this question directly in image and video generation, arguing that rewards should be modeled in latent space rather than inherited from pixel-space VLM evaluators. Learning rewards directly on noisy latent states has been shown to better align with latent-space generation and reduce the overhead of external evaluation [22, 27]. Relatedly, LikePhys [41] evaluates intuitive physics understanding in video diffusion models by comparing denoising losses on curated valid–invalid simulator-generated pairs and summarizing the preference with a Plausibility Preference Error (PPE) metric. In contrast, Proprio does not train a separate reward model, does not learn from pairs of diffusion models, and does not require curated valid–invalid physics pairs. Instead, it directly reuses the frozen generator's own residual under controlled perturbations as a training-free inference-time self-scoring signal for ranking and refinement.

H Broader and societal impact.

This work aims to improve the physical plausibility of generated videos, which may benefit applications such as robotics, simulation, and world modeling. More physically consistent video generation could make generative models more useful for studying agent behavior, planning, training embodied systems, and building controllable environments for decision-making and reasoning. At the same time, improving the realism of synthetic video also carries risks. More plausible generated videos may be harder to distinguish from real footage and could be misused for deception, misinformation, or other forms of misleading synthetic media. While we acknowledge these risks, our work is intended purely for academic research, and any examples produced in this work will be clearly labeled as synthetic. We hope this research contributes to safer and more controllable generative systems, while encouraging continued work on safeguards, transparency, and responsible deployment.

I Additional qualitative results

Figures 7 and 8 as well as figures 9 and 10 show additional qualitative results for T2V and I2V settings respectively. We compare baseline performance on TurboWan2.2 with the optimized version highlighting the improvement in physics plausibility of the generations. We summarize the prompts used with the videos above the figures.

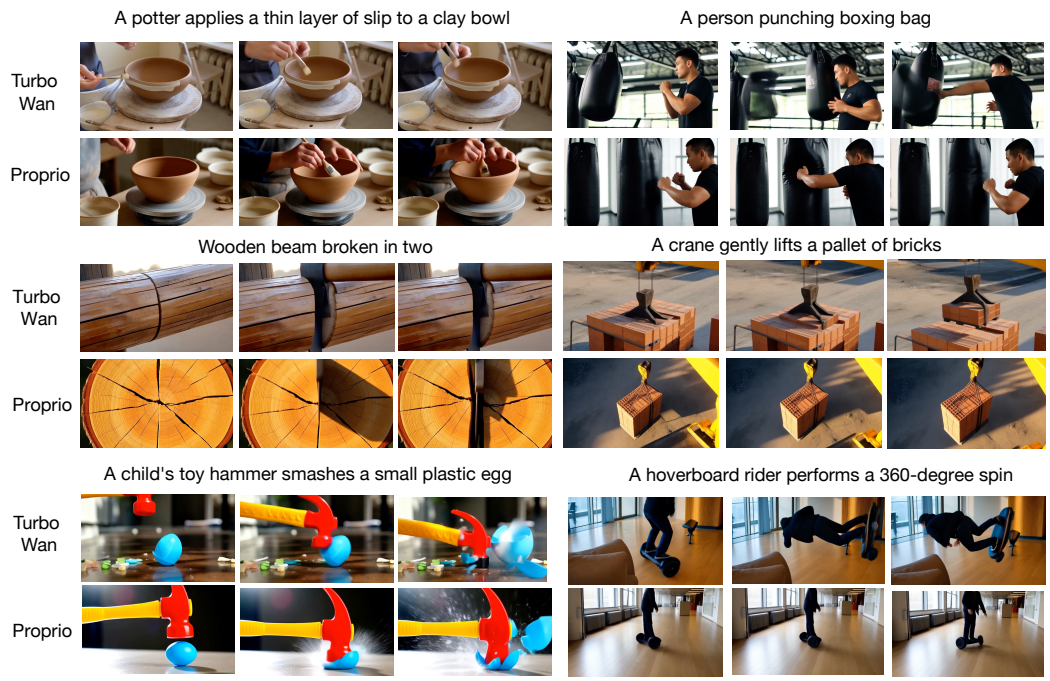


Figure 7: Overall qualitative results comparing our Proprio method against baseline TurboWan2.2. Visual examples from Proprio showing that our method produces more physically plausible generations for the T2V setting.

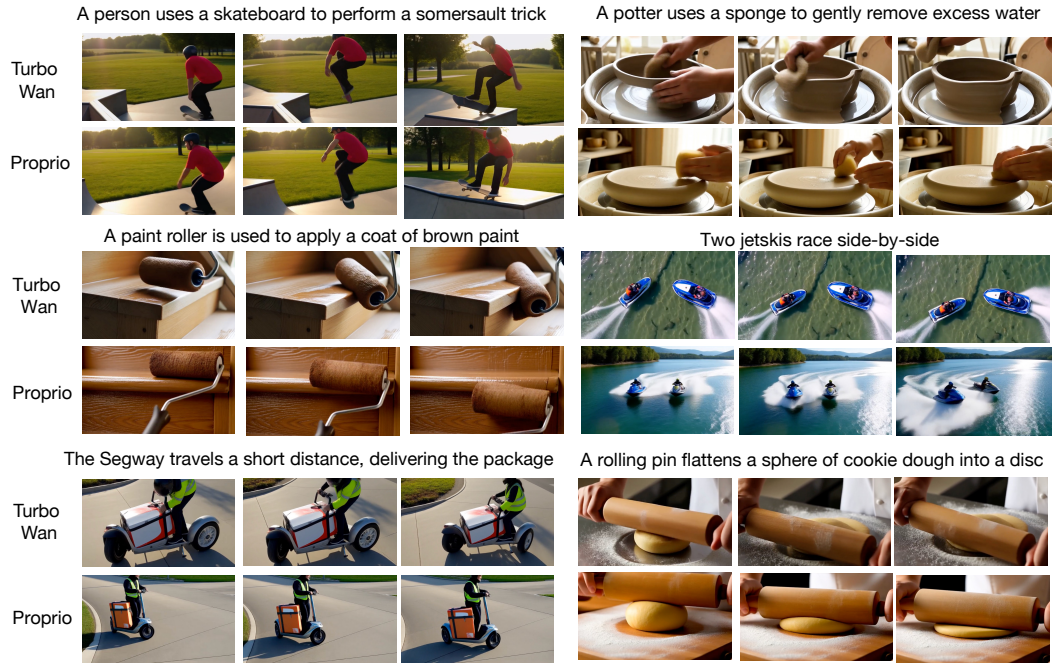


Figure 8: Overall qualitative results comparing our Proprio method against baseline TurboWan2.2. Visual examples from Proprio showing that our method produces more physically plausible generations for the T2V setting.

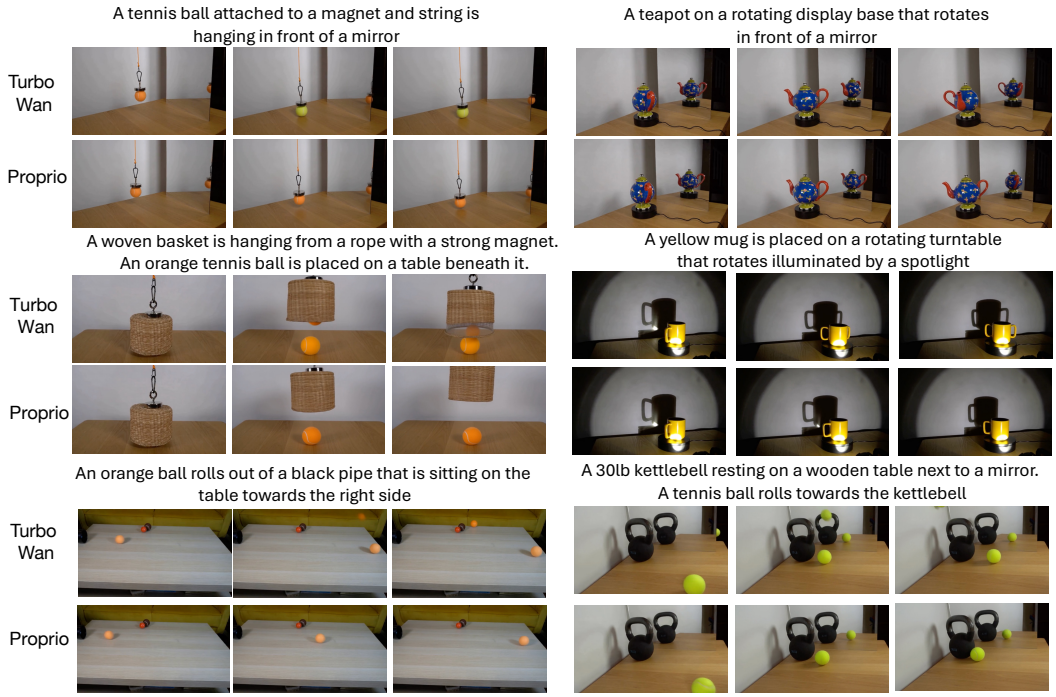


Figure 9: Overall qualitative results comparing our Proprio method against baseline TurboWan2.2. Visual examples from Proprio showing that our method produces more physically plausible generations for the I2V setting.

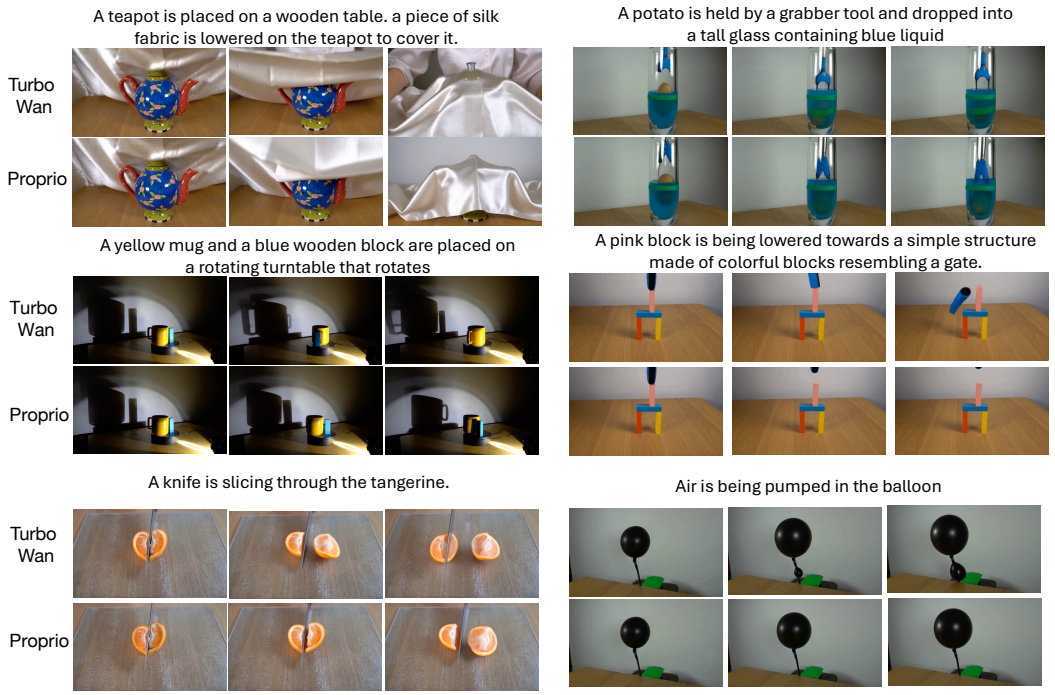


Figure 10: Overall qualitative results comparing our Proprio method against baseline TurboWan2.2. Visual examples from Proprio showing that our method produces more physically plausible generations for the I2V setting.

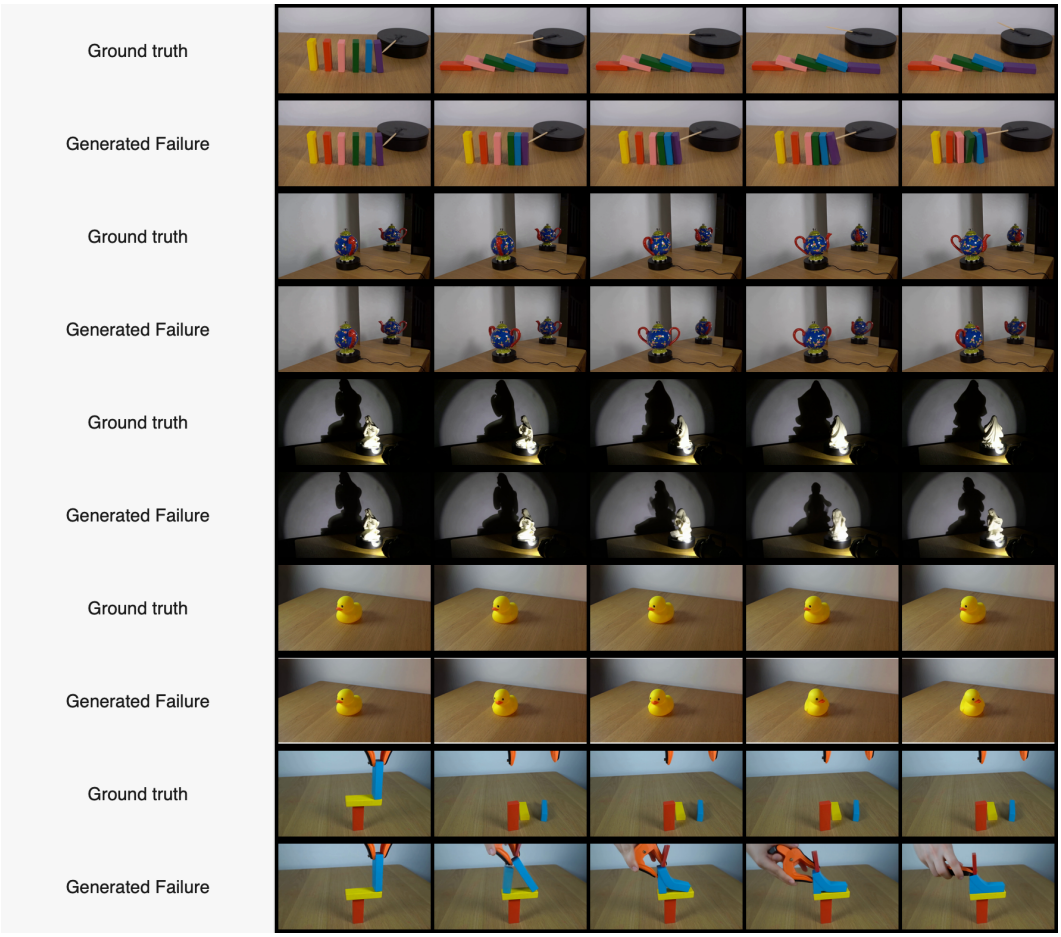


Figure 11: **Correctly preferred diagnostic examples.** Each pair shows a ground-truth Physics-IQ video and a generated Turbo Wan2.2 sample with a visible physical failure. These examples are drawn from the subset where the dynamic Proprio self-score correctly prefers the ground-truth video, illustrating that the generator-native residual can favor natural video dynamics over physically implausible generations.

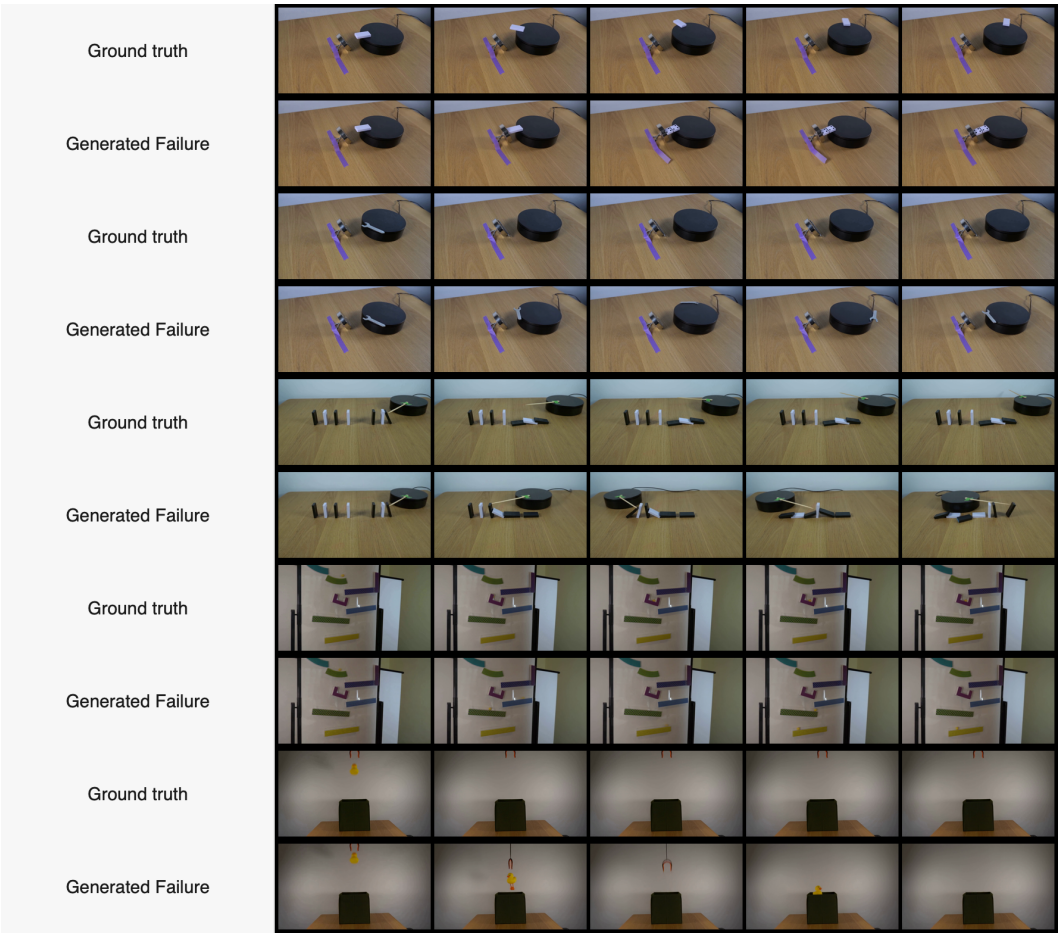


Figure 12: **Correctly preferred diagnostic examples.** Each pair shows a ground-truth Physics-IQ video and a generated Turbo Wan2.2 sample with a visible physical failure. These examples are drawn from the subset where the Proprio self-score correctly prefers the ground-truth video, illustrating that the generator-native residual can favor natural video dynamics over its own physically implausible generations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state and contributions that are tested and proven through empirical evidence in the experiments 4 and ablation 5 sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations as well as our future work in the conclusion section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include any theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully provide all settings for the baselines and experimental settings in section 4 and in the appendix E for reproducibility. We also will open-source our generation results and our code for easier reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be open sourced for reproducibility. All benchmarks used are already publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We use publicly available benchmarks and also provide in our code the json files with exact prompts and images used for all experiments. We provide a detailed section of implementation details and will open source the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide results across multiple seeds in A.6

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the resource details in the implementation details in section 4 and provide a comparative analysis of inference time in appendix A.3 and section 4.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper abides by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a discussion over the societal impact in Appendix H

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All benchmarks and models used are open source. We will add in our public code base proper licensing and disclaimer that it's for academic use only.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We carefully cited previous methods and all models and benchmarking datasets used are publicly available and eligible for academic use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code base will be properly documented upon publication for transparency and easier reproducibility. Otherwise, we do not propose any new datasets or architectures and our method is training-free on open source models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper includes human evaluation study, we present full instructions and screenshots in Appendix A.10.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We only have a human evaluation study where people volunteer to provide their opinion by choosing their preferred video. We do not save any information or emails.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.