WHAT ABOUT GRAVITY IN VIDEO GENERATION? POST-TRAINING NEWTON'S LAWS WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

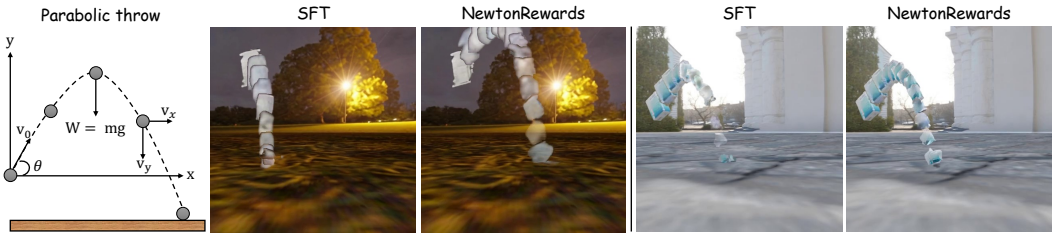

Figure 1: `NewtonRewards` enforces physical laws in video generation. Shown is a parabolic throw scenario from our NewtonBench-60K dataset. Baseline supervised fine-tuning (SFT) produces implausible motion violating Newtonian dynamics. Our `NewtonRewards` post-training restores parabolic trajectories that follow constant-acceleration behavior predicted by physics.

ABSTRACT

Video generators synthesize visually compelling clips, yet often violate basic physical laws–objects float or collide inconsistently, accelerations drift–revealing a gap between visual and physical realism. We propose `NewtonRewards`, the first physics-grounded post-training framework for video generation based on *verifiable rewards*. Instead of relying on human or VLM feedback, `NewtonRewards` extracts *measurable proxies* from generated videos using frozen utility models: optical flow serves as a proxy for velocity, while high-level visual features serve as a proxy for physical properties. These proxies enable explicit enforcement of Newtonian structure through two complementary rewards: a Newtonian kinematic constraint enforcing constant acceleration dynamics, and a physical properties conservation reward preventing trivial, degenerate solutions. We evaluate `NewtonRewards` on five Newtonian Motion Primitives (free fall, horizontal/parabolic throw, and ramp sliding down/up) using OpenSora and Wan 2.2 5B on our newly constructed large-scale benchmark, `NewtonBench-60K`. Across both models in visual and physics metrics, `NewtonRewards` consistently improves physical plausibility, motion smoothness, and temporal coherence over prior post-training methods. It further maintains strong performance across all primitives and under out-of-distribution shifts in height, speed, and friction. Our results show that physics-grounded verifiable rewards offer a scalable path toward physics-aware video generation.

1 INTRODUCTION

Gravity is everywhere. From the fall of an apple to the motion of celestial bodies, physical laws govern how objects move, interact, and persist through time. Yet, in the rapidly advancing field of video generation, such fundamental principles are largely absent (Motamed et al., 2025b; Bansal et al., 2025; Kang et al., 2025; Kong et al., 2024; Zhang et al., 2025a; Liu et al., 2025). Recent generative models can produce visually stunning videos from text (Zheng et al., 2024; Yang et al., 2024b; HaCohen et al., 2024; Gao et al., 2025; Wan et al., 2025; Le et al., 2026b), images (Blattmann et al., 2023a; HaCohen et al., 2024; Gao et al., 2025; Wan et al., 2025; Le et al., 2025; 2026a), or latent trajectories (Courant et al., 2024), but often exist in worlds unbound by physics: worlds where objects float (Yuan et al., 2025), collisions resolve unrealistically (Bansal et al., 2025), and motion unfolds without cause (Zhang et al., 2025a). These violations of basic laws, such as Newton's laws of motion, highlight a lack of *physical realism* in video generation.

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

Embedding physical plausibility into video generation is more than an aesthetic choice; it is a necessity. In applications, such as game environments and realistic worlds (Wang et al., 2025c; Bruce et al., 2024; Kanervisto et al., 2025; Yu et al., 2025), autonomous driving (Hu et al., 2023; Russell et al., 2025; Agarwal et al., 2025; Feng et al., 2025) and robotic control (Yang et al., 2023; Bar et al., 2025; Li et al., 2025a), generated videos serve as data for perception, reasoning, and action (Wu et al., 2025). In these cases, non-physical dynamics can lead to inconsistent learning signals, unrealistic affordances, and failure to generalize. A model that understands that "objects fall down" or that "collisions change velocity" produces not just more believable motion, but also a better world model.

Recent trends embed physical plausibility into video generation. These range from methods that fine-tune diffusion models using textual or feedback-based supervision from LLMs or VLMs (Lian et al., 2023; Zhang et al., 2025b; Hao et al., 2025; Yang et al., 2025; Motamed et al., 2025a; Xue et al., 2025) or align them with self-supervised video representations (Zhang et al., 2025c; Garrido et al., 2025), to approaches that incorporate physical simulators or 3D representations as motion or geometry priors (Xie et al., 2025; Wang et al., 2025a; Zhao et al., 2025; Li et al., 2025b; Liu et al., 2024; Yuan et al., 2026; Lv et al., 2024; Hsu et al., 2024), and to those that rely on physics-rich datasets or post-training signals from real or synthetic videos (Li et al., 2025b; Wang et al., 2025b; Chen et al., 2025; Ji et al., 2025). Visual similarity and plausible-looking motion do not, by themselves, establish whether a generated trajectory satisfies a physical constraint. This motivates a post-training signal that evaluates motion through explicit, measurable rules, and verifiable rewards (Lambert et al., 2024; Wei, 2025; Shao et al., 2024; Guo et al., 2025; Mroueh, 2025; Wen et al., 2025) for video dynamics.

We introduce `NewtonRewards`, a physics-grounded post-training framework that turns Newtonian motion constraints into differentiable, verifiable rewards. Our key idea is to evaluate governing relations on *measurable proxies* extracted from generated videos. Frozen utility models provide optical flow as a proxy for image-plane velocity and visual features as a proxy for physical properties. Rather than matching a reference flow field, our kinematic reward penalizes deviations from a constant-acceleration constraint shared by the five Newtonian Motion Primitives (NMPs) considered here. Crucially, this constraint alone admits trivial solutions: eliminating motion or making objects disappear can also reduce its residual. A complementary physical properties conservation reward anchors generated features to those of simulated reference videos, discouraging these degeneracies. Together, the rewards couple explicit supervision of motion with preservation of object consistency, providing a direct training signal beyond visual similarity alone.

We evaluate `NewtonRewards` on our simulated **NewtonBench-60K** across *free fall, horizontal and parabolic throws*, and *ramp sliding down/up*. On OpenSora (Zheng et al., 2024) and Wan 2.2 5B (Wan et al., 2025), it improves all five visual and physics metrics over SFT, with average relative gains of 9.75% and 13.56%, respectively. Ablations show that feature alignment alone recovers only a small fraction of these gains, highlighting the importance of the explicit kinematic constraint. Further OpenSora experiments demonstrate improvements under out-of-distribution shifts.

Our contributions are: (1) We introduce `NewtonRewards`, a physics-grounded post-training framework that turns Newtonian motion constraints into differentiable, rule-based rewards on measurable video proxies. (2) We combine an optical-flow-based kinematic residual with a reference-feature physical properties conservation reward, coupling explicit motion constraints with object consistency to mitigate degenerate solutions. (3) We simulate a controlled, large-scale dataset and benchmark specifically designed to evaluate dynamic motion realism and physical consistency in video generation. We will release our simulation/training code, `NewtonBench-60K`, and models to the community. (4) Experiments on OpenSora and Wan 2.2 5B show that `NewtonRewards` consistently improves across both visual and physics metrics, with more physically faithful videos.

Position. We posit that our methodology is general. *Given a measurable proxy of a variable in a physical law, the same steps can be followed to come up with verifiable rewards appropriate* for that law. We hope that this general framework can pave the way for future research in this area.

Scope. Our goal is to introduce a paradigm of *measurable proxies* and *verifiable rewards* for video post-training, rather than to claim we cover all physics. As the first work to introduce verifiable rewards for post-training video generation, we start from Newton's laws as a concrete instantiation, laying the foundation for extensions. Yet, our framework is not limited to these primitives: deformable objects can be modeled with particle simulators, and fluid dynamics with Navier–Stokes constraints.

Table 1: Natural motion primitives in `NewtonBench-60K`.

NMP	Description	Example
Free fall	Spawn heights $\sim [0.5, 1.5]$ m; zero initial velocity.	
Horizontal throw	Horizontal speed $v_0 \in [2, 6]$ m/s; pitch 0°.	
Parabolic throw	Speed $v_0 \in [2, 6]$ m/s; launch angle $\theta \in [15^\circ, 75^\circ]$.	
Ramp slide down	Ramp angle $\theta \in [15^\circ, 45^\circ]$, kinetic friction $(\mu_{\text{ramp}}, \mu_{\text{obj}}) \approx 0.06$; released from rest near the crest.	
Ramp slide up	Same ramp; initial uphill velocity $v_0 \in [3, 4]$ m/s along the ramp tangent.	

2 Newtonian Motion Primitives (NMPs)

We formulate post-training of video diffusion through classical mechanics, grounded in Newton's three laws of motion (Section 2.1). In our setting, an object observed in a video sequence undergoes motion determined by external forces acting upon it. Thus, we identify five canonical *Newtonian Motion Primitives (NMPs)*: *free fall*, *horizontal throw*, *parabolic throw*, *ramp sliding down*, and *ramp sliding up*. Each primitive corresponds to a distinct combination of forces and produces a characteristic pattern of constant acceleration in the image plane (Section 2.2).

2.1 Background: Newton's Laws of Motion

Let $\mathbf{F}_{\text{net}}$ denote the net resultant force acting on the object, $\mathbf{v} = (\mathbf{v}_x, \mathbf{v}_y)$ its velocity projected onto the image plane, and $\mathbf{a} = \dot{\mathbf{v}}$ its corresponding acceleration.

Newton's three laws: **(1)** *Inertia*: $\mathbf{F}_{\text{net}}{=}\mathbf{0} \Rightarrow$ uniform motion; in free fall, horizontal, and parabolic throws only gravity acts, so the horizontal net force is zero ($F_{\text{net},x}{=}0$), which yields $\mathbf{a}_x{=}0$ and constant $\mathbf{v}_x$. **(2)** *Acceleration*: $\mathbf{a} = \mathbf{F}_{\text{net}}/m$, links motion to forces and mass. **(3)** *Action–Reaction*: two interacting bodies exert equal, opposite forces; on a ramp, the surface exerts a normal force balancing contact, together with a tangential friction force opposing motion.

2.2 Forces and Accelerations of NMPs

Let the world coordinate system be (X, Y, Z), with gravity acting as $\mathbf{g} = (0, -g, 0)$, and let (x, y) denote the projection in the image plane.

Under a pinhole camera model with focal length f and scene depth Z, the image-plane coordinates scale with depth, giving the approximation $x = (f/Z)X$. For short intervals, this scale is nearly constant, allowing analysis in image space. [1]

(1–3) NMP-F/TH/TP: Free Fall, Horizontal and Parabolic Throw. Object motion under uniform gravity $\mathbf{g}$ and no other external forces follows Newton's Second Law: $\mathbf{a}_x = 0$ and $\mathbf{a}_y = -g$, corresponding to constant downward acceleration. Different initial conditions yield special cases: zero initial velocity for free fall (NMP-F), nonzero horizontal velocity for horizontal throw (NMP-TH), and arbitrary initial velocity $(\mathbf{v}_{0x}, \mathbf{v}_{0y})$ for parabolic throw (NMP-TP).

[1] We assume: the image's vertical axis is aligned with gravity's direction. For tilted cameras, projecting $\mathbf{g}$ onto image plane yields constant apparent acceleration under mild assumptions: weak perspective (small depth variation) and a static camera.

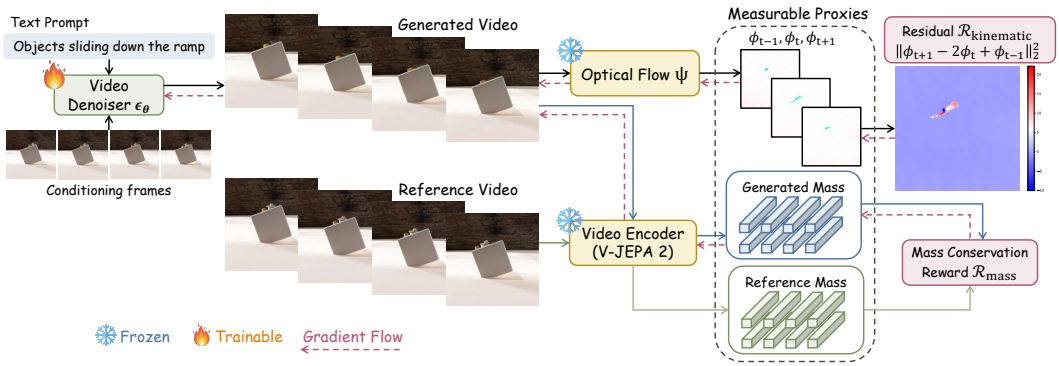

Figure 2: **Physics-Grounded Video Post-Training Pipeline.** Our method improves a pre-trained video generator by using physics-based rewards. Utility models (optical flow Ψ and V-JEPA 2) process the generated video to compute measurable proxies, from which kinematic and physical properties conservation rewards are derived to enforce explicit physics constraints.

(4-5) NMP-RD and NMP-RU: Ramp Sliding with Friction. For a ramp inclined at angle θ with kinetic friction coefficient μ_k, decomposing gravity along the downhill ramp tangent $\hat{\mathbf{s}} = (t_x, t_y)$, with friction opposing the motion, yields a mass-independent tangential acceleration $\mathbf{a}_s = g(\sin\theta \mp \mu_k \cos\theta)$ for sliding down/up (when sliding up, gravity and friction both act downhill and decelerate the object), constant within each sliding phase and projected onto the image plane as $\mathbf{a} = (\mathbf{a}_s t_x, \mathbf{a}_s t_y)$. Force decomposition in Appendix B.4.

3 METHOD: NewtonRewards

NewtonRewards post-trains video generators to follow physical laws. The main challenge lies in constructing reward losses that enforce *physical constraints*.

Given that the physics quantities are not directly measurable from generated videos, our core idea is to leverage *measurable proxies* extracted from generated videos to construct these *rewards*.

NewtonRewards has two components (Figure 2): (i) computing measurable proxies from model outputs (Section 3.1), and (ii) defining reward functions that quantify adherence to physics constraints; Section 3.2 details how these proxies are used to construct reward functions that enforce physical constraints. These reward signals are then used to fine-tune video generators, so that generated sequences not only appear realistic but also follow physical laws.

3.1 MEASURABLE PROXIES

Let $G_\theta(\epsilon, c)$ denote a video generator with parameters θ, producing a video $V = G_\theta(\epsilon, c)$ given initial latent noise ϵ and condition c. We define a measurable-proxy extractor M which maps the generated video V to a set of differentiable, physically meaningful quantities $M(V)$ (e.g., per-pixel displacement fields, object velocities, or visual features) that can serve as measurable proxies for physics quantities. In this work, we extract optical flow fields as proxies for velocity (Section 3.1.1) and visual appearance embeddings as proxies for physical properties (Section 3.2.1). These allow us to formulate physics-grounded constraints, such as constant-acceleration residuals and physical properties-conservation consistency, that provide verifiable reward signals for fine-tuning the generator.

3.1.1 OPTICAL FLOW AS VELOCITY PROXY.

In real-world videos we cannot directly observe $\mathbf{v}_t$. We instead employ an optical flow model $M_{OF} = \mathbf{\Psi}$ to estimate the per-frame displacement field from consecutive video frames V_t, V_{t+1}:

$$\boldsymbol{\phi}_t = \mathbf{\Psi}(V_t, V_{t+1}) \quad , \tag{1}$$

where $\boldsymbol{\phi}_t = (\phi_t^x, \phi_t^y)$ denotes predicted optical flow (in pixels/frame). The image-plane velocity is approximated as:

$$\mathbf{v}_t \approx \boldsymbol{\phi}_t / \Delta t, \quad \Delta t = 1/\text{FPS}. \tag{2}$$

3.1.2 VISUAL FEATURES AS PHYSICAL PROPERTIES PROXY.

Newton's Second Law states that acceleration is inversely proportional to physical mass property under fixed external force, implying that heavier objects exhibit smoother, slower changes in motion. Though absolute physical properties are not directly observable in videos, visual appearance and texture cues often correlate with object identity, shape, material, and thus effective physical properties. We define a *physical properties proxy* based on high-level visual representations extracted from a pre-trained self-supervised video encoder, whose features are known to capture intuitive physical properties (Garrido et al., 2025). Let $\mathbf{z}_t = M_{\text{physical properties}}(V_t)$ denote the per-frame per-patch feature embedding from the encoder, capturing consistent object-level visual information over time. This space provides a differentiable, semantically aligned measure of visual physical properties, which can be compared across space and time between simulated and generated videos.

3.2 PHYSICS CONSTRAINTS AND REWARDS

Let $\{C_j\}_{j=1}^{J}$ denote a set of physical constraints or laws that should be satisfied by the extracted proxy quantities $m_j = M_j(V)$. We construct a physics penalty

$$\mathcal{L}_{\text{phys}} = \sum_{j=1}^{J} \lambda_j \cdot \ell(C_j[M_j(V)]) \quad , \tag{3}$$

where $\ell(\cdot)$ is a penalty or norm (e.g., squared error or hinge loss) and λ_j are weighting coefficients.

3.2.1 DISCRETE CONSTANT-ACCELERATION CONSTRAINT.

For all primitives above, the image-plane accelerations $(\mathbf{a}_x, \mathbf{a}_y)$ remain constant throughout the motion. Discretizing the kinematic relation $\mathbf{v}_{t+1} = \mathbf{v}_t + \mathbf{a}\,\Delta t$ and $\mathbf{v}_t = \mathbf{v}_{t-1} + \mathbf{a}\,\Delta t$ yields

$$\mathbf{v}_{t+1} - 2\mathbf{v}_t + \mathbf{v}_{t-1} = \mathbf{0} \quad . \tag{4}$$

This discrete second-derivative constraint enforces the constant acceleration implied by Newton's Second Law. Substituting the proxy in Equation 2 into Equation 4, we obtain the unified residual for all NMPs (see Section 2.2):

$$\boldsymbol{\phi}_{t+1} - 2\,\boldsymbol{\phi}_t + \boldsymbol{\phi}_{t-1} \approx \mathbf{0} \quad . \tag{5}$$

Proposition 1 (Newtonian Kinematic Constraint)**.** *For an object governed by time-invariant external forces, the discrete second-order derivative of its optical-flow field predicted by $\boldsymbol{\Psi}$ vanishes:*

$$\mathcal{R}_{\textit{kinematic}} = \|\boldsymbol{\phi}_{t+1} - 2\,\boldsymbol{\phi}_t + \boldsymbol{\phi}_{t-1}\|_2^2 \approx \mathbf{0} \quad . \tag{6}$$

This is the optical-flow realization of Newton's Second Law in the video domain, enforcing constant acceleration across all five Newtonian Motion Primitives.

Equation equation 6 thus defines our *constant-acceleration residual* $\mathcal{R}_{\text{kinematic}}$, capturing the essence of Newtonian dynamics in a differentiable, video-aligned form.

3.2.2 PHYSICAL PROPERTIES CONSERVATION VIA VISUAL FEATURES.

Given reference embeddings $\mathbf{z}_t^{\text{sim}}$ from physically simulated videos and $\mathbf{z}_t^{\text{gen}}$ from generated videos, the high-level feature similarity constraint encourages physical properties preservation, so that the generator maintains consistent object visual appearance–hence consistent inferred physical properties–throughout the sequence and across domains. The physical properties conservation residual

$$\mathcal{R}_{\text{physical-properties}} = \frac{1}{T}\sum_{t=0}^{T-1} \left\|\mathbf{z}_t^{\text{gen}} - \mathbf{z}_t^{\text{sim}}\right\|_2^2 \quad , \tag{7}$$

penalizes deviations between generated and simulated visual features. Minimizing $\mathcal{R}_{\text{physical-properties}}$ encourages the generator to produce motions and appearances that obey physical properties-dependent dynamics implied by Newton's laws, complementing the kinematic constraint in Equation 6.

3.2.3 POST-TRAINING OBJECTIVE.

Our final objective combines a Newtonian kinematic constraint (Prop. 1) and a physical properties-matching term (Eq. 7):

$$\mathcal{L}_{\text{phys}} = \lambda_{\text{kinematic}}\,\mathcal{R}_{\text{kinematic}} + \lambda_{\text{physical-properties}}\,\mathcal{R}_{\text{physical-properties}} \quad . \tag{8}$$

As the background is static and the camera is fixed in our dataset, the optical flow ϕ directly captures object motion, letting us compute the loss over the entire frame. This enforces Newtonian consistency without requiring explicit acceleration/depth supervision, unifying all five primitives (Section 2.2) under a single principle: under constant external forces, image-plane accelerations remain constant.

4 NEWTONBENCH-60K

We introduce **NewtonBench-60K**, a controlled, large-scale dataset and benchmark designed to isolate and evaluate five *Newtonian Motion Primitives (NMPs)*: *free fall*, *horizontal throw*, *parabolic throw*, *sliding down a ramp with friction*, and *sliding up a ramp with an uphill initial velocity*, each visualized in Table 1. The corpus comprises **50K** simulated training videos (**10K** per NMP) and a **10K** held-out benchmark with **2K** videos per NMP, evenly split into *In-Distribution* (ID) and *Out-Of-Distribution* (OOD) subsets. In comparison, PISA (Li et al., 2025b) focuses solely on free-fall and does not capture a broader range of Newtonian dynamics.

4.1 SIMULATION PIPELINE

We use *Kubric* (Greff et al., 2022) for scene orchestration, *PyBullet* for rigid-body dynamics, and *Blender* for rendering, with gravity $\mathbf{g} = (0, 0, -9.81)$. Videos are rendered at 512×512, 32 frames at 16 fps with HDRI lighting and a fixed side-view camera (for unambiguous 2D motion analysis). Objects are sampled from GSO (Downs et al., 2022) and backgrounds from HDRI (Greff et al., 2022) with disjoint train/test pools. Each NMP is generated by sampling a small set of physically meaningful parameters (Table 1). For evaluation, we use renderer instance masks for ground truth and SAM2 (Ravi et al., 2025) masks (guided by conditioning frames and object prompts) for generated videos; centroids $\mathbf{c}^{\text{gen}}$ are computed from these masks. Full pipeline details (per-clip outputs, ramp construction, etc.) are in Appendix B.

4.2 BENCHMARK PROTOCOL: ID & OOD

For each NMP, we synthesize **1K ID** and **1K OOD** videos. ID parameter ranges mirror training (e.g., fall height $[0.5, 1.5]$ m; throw speed $[2, 6]$ m/s; ramp angle $[15°, 45°]$). OOD ranges deliberately hold out disjoint bands to stress generalization: *lower* horizontal throws (e.g., $[1.7, 2.0]$ m/s), *lower* parabolic throws (e.g., $[1.7, 2.0]$ m/s), *extreme* parabolic angles (e.g., $(75°, 90°]$), and *steeper/shallower* ramps (e.g., $(45°, 60°]$ with *faster* sliding up $[4.0, 5.0]$ m/s). We optionally perturb friction by $\pm25\%$ in OOD to decouple appearance from dynamics.

4.3 EVALUATION METRICS

We evaluate generated videos with two **physics metrics**, $\text{RMSE}_{\mathbf{v}}$ and $\text{RMSE}_{\mathbf{a}}$, computed as the root-mean-square error between generated and ground-truth object-centroid first and second discrete time derivatives, capturing first- and second-order Newtonian consistency, and three **visual metrics** (following PISA (Li et al., 2025b)): trajectory position error (centroid ℓ_2), Chamfer Distance (CD) between per-frame binary masks, and per-frame Intersection over Union (IoU). All metrics are computed per object and averaged across videos; full equations are in Appendix B.6.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTING.

We fine-tune our base video diffusion models Open-Sora v1.2 (Zheng et al., 2024) and Wan 2.2 5B (Wan et al., 2025) on our `NewtonBench-60K` dataset comprising 50K simulated videos across

Table 2: Comparison of post-training strategies on OpenSora and Wan 2.2 5B SFT. PISA ORO variants follow (Li et al., 2025b); the w/o rows ablate the kinematic and physical-property reward components of `NewtonRewards`. Visual metrics (L2, CD, IoU) capture pixel alignment and shape agreement; physics metrics ($\text{RMSE}_\mathbf{v}$, $\text{RMSE}_\mathbf{a}$) capture physical plausibility in motion. Avg. Change is the mean relative improvement over all five metrics vs. the SFT baseline. Best result in **bold**.

	Visual metrics			Physics metrics		
Method	**L2** ↓	**CD** ↓	**IoU** ↑	$\textbf{RMSE}_\mathbf{v}$ ↓	$\textbf{RMSE}_\mathbf{a}$ ↓	**Avg. Change** ↑
OpenSora (SFT)	0.110	0.316	0.110	0.279	3.32	–
+ PISA Optical Flow	0.104	0.296	0.118	0.280	2.72	+7.6%
+ PISA Depth Map	0.108	0.311	0.115	0.288	3.47	+0.4%
+ PISA Segmentation	0.110	0.318	0.114	0.280	3.29	+0.7%
+ `NewtonRewards`						
w/o kinematic	0.111	0.320	0.115	0.279	3.33	+0.3%
w/o phys. props.	0.106	0.299	0.117	0.274	**2.54**	+8.1%
+ `NewtonRewards` full	**0.096**	**0.293**	**0.127**	**0.263**	3.04	**+9.8%**
Wan 2.2 5B (SFT)	0.065	0.247	0.332	0.377	8.70	–
+ PISA Optical Flow	0.066	0.215	0.343	0.382	7.66	+5.3%
+ PISA Depth Map	0.074	0.224	0.317	0.381	8.72	−2.0%
+ PISA Segmentation	0.064	**0.193**	0.340	0.300	8.79	+9.1%
+ `NewtonRewards`						
w/o kinematic	0.065	0.255	0.335	0.370	8.29	+0.9%
w/o phys. props.	**0.062**	0.217	0.346	0.313	7.15	+11.2%
+ `NewtonRewards` full	0.063	0.216	**0.347**	**0.293**	**6.56**	**+13.6%**

five NMPs, and fine-tune them to accept both text and the first 4 frames of a video as conditions, respectively. Both Supervised Fine-Tuning (SFT) and post-training operate on 32-frame clips at 16 fps, consistent with the dataset specification. Training is performed on 8×NVIDIA H100 (80GB) GPUs with a batch size of 1 and gradient accumulation of 32. The learning rate is set to 1×10^{-4} for SFT and 1×10^{-5} for post-training. In the post-training objective (Eq. 8), we use equal weighting $\lambda_{\text{kinematic}} = \lambda_{\text{physical-properties}} = 1.0$. We employ the RAFT (Teed & Deng, 2020) optical-flow model to compute motion fields and the V-JEPA 2 (Assran et al., 2025) encoder to extract visual features for physical properties alignment. Evaluation follows the `NewtonBench-60K` protocol.

5.2 Experimental Results

Comparison with State of The Art. We compare `NewtonRewards` framework with three post-training strategies adapted from PISA (Li et al., 2025b): *Optical Flow Reward*, *Depth Reward*, and *Segmentation Reward*. For each model, all methods are fine-tuned from the same SFT baseline under identical settings on `NewtonBench-60K`. In PISA, each reward measures the similarity between the generated video and its simulated ground truth: RAFT (Teed & Deng, 2020) computes optical flow fields and minimizes their discrepancy, Depth-Anything-V2 (Yang et al., 2024a) aligns predicted and true depth maps, and SAM2 (Ravi et al., 2025) provides object masks for IoU-based supervision. Table 2 reports the results for OpenSora and Wan 2.2 5B. We observe that visual feature-based rewards improve appearance metrics, for instance, Depth and Optical Flow rewards slightly enhance L2, CD, and IoU; our physical properties reward improves IoU. However, they do not guarantee physically consistent motion–reflected in higher velocity and/or acceleration errors ($\text{RMSE}_\mathbf{v}$, $\text{RMSE}_\mathbf{a}$).

Finding 1. *Visual feature alignment improves perceptual and spatial fidelity but fails to enforce adherence to physical laws of motion.*

On OpenSora in Table 2, PISA Depth shows small spatial gains (+1–4%) but degrades temporal consistency (−3% in $\text{RMSE}_\mathbf{v}$, −4% in $\text{RMSE}_\mathbf{a}$). Post-training only with our physical properties conservation reward shows a similar trend (`NewtonRewards` w/o kinematic). Segmentation and optical flow rewards slightly worsen velocity measures, suggesting that frame-level feature alignment alone cannot capture Newtonian dynamics. On Wan, Segmentation improves CD and velocity error by +21.72% and +20.33%, respectively, but worsens acceleration error by −1.08%; Depth yields an

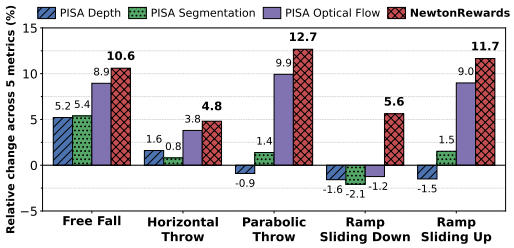

Figure 3: **Relative performance change across Newtonian Motion Primitives.** Percentage gains over the SFT baseline across all five NMPs. `NewtonRewards` consistently improves across primitives, while PISA rewards show task-dependent gains and regressions.

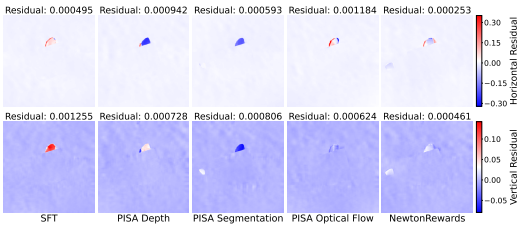

Figure 4: Mean horizontal and vertical residuals $\phi_{t+1} - 2\,\phi_t + \phi_{t-1}$ on OpenSora. Lower magnitude indicates closer adherence to constant-acceleration dynamics. `NewtonRewards` produces the smallest residuals, while SFT and PISA variants show larger deviations.

average change of -2.01%. In contrast, `NewtonRewards` achieves consistent improvements across all five metrics on both OpenSora (average $+9.75\%$) and Wan (average $+13.56\%$), demonstrating that enforcing self-consistent kinematic and physical properties constraints provides a stronger inductive bias for physically grounded and temporally coherent video generation. On Wan, `NewtonRewards` achieves the best IoU, velocity, and acceleration results, reducing velocity and acceleration errors by $+22.25\%$ and $+24.65\%$, respectively.

Evaluation Across Newtonian Motion Primitives. Figure 3 shows the relative performance gain of each post-training strategy over the SFT baseline across all five NMPs. PISA rewards show inconsistent and task-dependent behavior: Depth is mildly better on free fall and horizontal throw, but worse on parabolic and ramp dynamics. Same with Segmentation: small gains on simple trajectories but clear regression on ramp-down and marginal improvement elsewhere. Optical Flow is highly unstable, with large swings across primitives, including strong regression on ramp-down despite large gains on free fall and parabolic throw. `NewtonRewards` provides *uniformly positive* gains across all primitives with largest gains on most challenging motions: **+12.7%** on parabolic throws and **+11.7%** on ramp-up, while outperforming all baselines on simple trajectories.

Qualitative Comparison. While PISA-based rewards grounded in visual similarity (Depth, Segmentation, Optical Flow) sometimes improve local appearance, they fail to enforce physically coherent motion. In contrast, our `NewtonRewards` yields more coherent trajectories and consistent object appearance. Representative Wan and OpenSora comparisons are in Appendix H, Figures 7 and 8. Together with the quantitative trends in Table 2 and Figure 4, these observations support that physics-grounded verifiable rewards promote perceptual fidelity and dynamic realism.

OOD Evaluation. `NewtonRewards` also generalizes to out-of-distribution settings (higher drops, faster throws, steeper ramps, and perturbed friction): it improves over the SFT baseline across all five metrics (average $+8.60\%$), including a $+9.79\%$ reduction in $\mathrm{RMSE}_{\mathbf{v}}$ and $+16.09\%$ in $\mathrm{RMSE}_{\mathbf{a}}$, despite never observing OOD dynamics during post-training. This shows that enforcing Newtonian kinematic and physical properties constraints yields models that extrapolate more reliably to unseen physical regimes. Full per-metric results are in Appendix C.

Constant-Acceleration Residual Analysis. To assess whether generated motions obey Newtonian kinematics, we compute the mean discrete second-order residual $\phi_{t+1} - 2\,\phi_t + \phi_{t-1}$ across all 32 frames of the ramp-sliding-down scenario (Appendix H, Figure 8). This residual is zero for ideal constant-acceleration motion and thus serves as a diagnostic of dynamical consistency. Figure 4 shows the horizontal (top) and vertical (bottom) residual fields. The SFT baseline and all PISA variants exhibit strong red/blue activations, indicating clear violations of the constant-acceleration constraint. Even methods using ground-truth visual signals (PISA Depth, Segmentation, Optical Flow) retain structured residuals, suggesting that pixel-level alignment does not ensure correct dynamics. In contrast, `NewtonRewards` produces smoother residual maps with minimal magnitude, with the lowest absolute residuals along both axes and trajectories that better follow constant-acceleration.

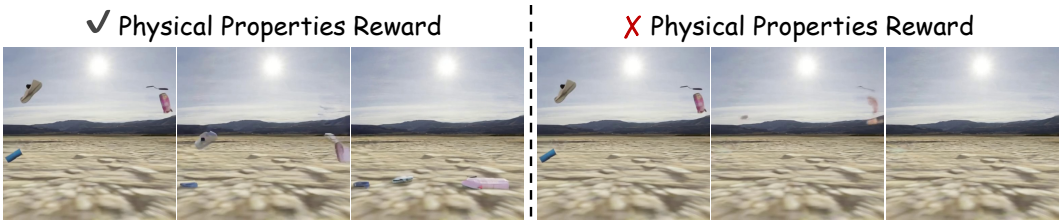

Figure 5: Videos generated with the physical properties reward (left) maintain consistent object persistence, while removing it (right) leads to degenerate behavior where objects vanish; an instance of reward hacking when optimizing only the kinematic residual.

5.3 Ablation Study

Newtonian Kinematic Residual Constraint. On OpenSora, ablating the discrete residual term largely removes the temporal regularization effect, leading to only marginal overall gain over SFT (+0.25%). Without enforcing constant acceleration, the model produces visually coherent but physically inconsistent motion: slightly better spatial fidelity (IoU +3.8%) yet degraded kinematic accuracy across L2, CD, velocity, and acceleration metrics, showing the importance of residual constraint for stabilising motion and aligning dynamics with Newtonian laws. On Wan, removing the kinematic constraint similarly reduces the average gain from +13.56% to +0.93%, with velocity and acceleration improvements falling to +1.75% and +4.76%, respectively.

Physical Properties Conservation Reward. Although the overall gain decreases when removing physical properties conservation (+8.1% vs. +9.75% on OpenSora; +11.16% vs. +13.56% on Wan), as seen in Table 2, this configuration only employs the Newtonian kinematic residual without additional regularization. On Wan, the full reward improves CD, IoU, velocity, and acceleration over this ablation, while the ablation yields a slightly lower L2 error. To understand its influence, we assess the role of physical properties conservation in stabilizing post-training.

Finding 2. *The physical properties conservation reward mitigates reward hacking that emerges when optimizing solely the kinematic residual, thereby avoiding trivial solutions.*

Without this constraint, the model converges to a degenerate, trivial solution that minimizes the residual by driving all velocities to zero ($\mathbf{v}_t = 0$), effectively causing object disappearance. By anchoring visual and feature-level physical properties consistency, the physical properties reward prevents this collapse and ensures stable, physically meaningful motion as in Figure 5, refer to Appendix G for quantitative analysis. We further analyze synthetic video pairs with different object masses. For each pair, we compute two differences: mass $\mathbf{m}_1 - \mathbf{m}_2$ and features (frozen V-JEPA 2 (Assran et al., 2025)) $\mathbf{z}_1 - \mathbf{z}_2$ and find a strong correlation between them (Pearson 0.7737, Spearman 0.7262). This indicates that V-JEPA 2 features implicitly represent mass, not only semantic alignment.

6 Conclusion

We introduced `NewtonRewards`, a general physics-grounded post-training framework that enforces Newtonian consistency in video generation by constructing verifiable rewards from measurable proxies. By leveraging OF and visual features as surrogates for velocity and physical properties, `NewtonRewards` enforces Newtonian dynamics through kinematic and physical properties-conservation constraints. Our method enables generators to obey constant-acceleration dynamics and maintain physical plausibility across diverse NMPs, with post-training gains on both OpenSora and Wan 2.2 5B. Our approach is inherently general. It unlocks a broader view of physics-grounded post-training: once a physical quantity can be estimated, generative models can be guided toward physically valid behavior through explicit, differentiable constraints.

AI USE STATEMENT

We used AI-assisted tools to support debugging, visualization, literature search, and language editing. These tools supported the presentation of the work and were not treated as independent evidence for scientific claims. The authors retain responsibility for verifying AI-assisted content, including references, validating the implementation and experimental results, and ensuring the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

Our experiments use simulated scenes and existing research datasets to study physical consistency in video generation. While improved physical realism can benefit scientific visualization and simulation, it may also increase the credibility of deceptive synthetic media. We emphasize responsible use, appropriate attribution of source assets, and clear disclosure of generated content. Our evaluation covers a limited set of Newtonian motions and does not establish general physical correctness or suitability for safety-critical applications.

REPRODUCIBILITY STATEMENT

The main paper and Appendix describe the dataset construction and splits, reward formulations, training settings, and evaluation metrics. Comparisons use matched initialization, architecture, training data, and conditioning within each backbone to isolate the effect of reward formulation. We identify the pretrained generators and frozen utility models used in our framework and provide implementation details and metric definitions to support reproduction and interpretation of the reported results.

REFERENCES

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

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Aritra Bhowmik, Denis Korzhenkov, Cees GM Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. *arXiv preprint arXiv:2510.19022*, 2025.

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Tim Brooks, Bill Peebles, Connor Homes, 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 technical report, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. *arXiv preprint arXiv:2508.10858*, 2025.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.T. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *ECCV*, 2024.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

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

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

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.

Yutong Hao, Chen Chen, Ajmal Saeed Mian, Chang Xu, and Daochang Liu. Enhancing physical plausibility in video generation by reasoning the implausibility. *arXiv preprint arXiv:2509.24702*, 2025.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022b.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

Hao-Yu Hsu, Zhi-Hao Lin, Albert Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically realistic video editing from natural language instructions. *arXiv preprint arXiv:2411.02394*, 2024.

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

Sihui Ji, Xi Chen, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Physmaster: Mastering physical representation for video generation via reinforcement learning. *arXiv preprint arXiv:2510.13809*, 2025.

Anssi Kanervisto, Dave Bignell, Linda Yilin Wen, Martin Grayson, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Tabish Rashid, Tim Pearce, Yuhan Cao, et al. World and human action models towards gameplay ideation. *Nature*, 638(8051):656–663, 2025.

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*, 2025.

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

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Minh-Quan Le, Gaurav Mittal, Tianjian Meng, A S M Iftekhar, Vishwas Suryanarayanan, Barun Patra, Dimitris Samaras, and Mei Chen. Hummingbird: High fidelity image generation via multimodal context alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6kPBThI6ZJ.

Minh-Quan Le, Armand Comas, Alexandros Lattas, Stylianos Moschoglou, Pedro Vélez, Amit Raj, Aaron Germuth, Thabo Beeler, Dimitris Samaras, and Di Qiu. Concurrent image understanding and generation: Self-correcting coupled markov jump processes. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=EzyXXt9tbh.

Minh-Quan Le, Gaurav Mittal, Cheng Zhao, David Gu, Dimitris Samaras, and Mei Chen. PISCES: Annotation-free text-to-video post-training via optimal transport-aligned rewards. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=wSfc8mDEjM.

Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pp. 34–50. Springer, 2022.

Chenhao Li, Andreas Krause, and Marco Hutter. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100*, 2025a.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025b.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023.

Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, Ajmal Mian, Mubarak Shah, and Chang Xu. Generative physical ai in vision: A survey. *arXiv preprint arXiv:2501.10928*, 2025.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *ECCV*, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility. *arXiv preprint arXiv:2510.07550*, 2025a.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025b.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. PHANTOM: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11185–11194, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shen_PHANTOM_Physics-Infused_Video_Generation_via_Joint_Modeling_of_Visual_and_CVPR_2026_paper.html.

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

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

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, 2025. URL https://arxiv.org/abs/2503.20314.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. *arXiv preprint arXiv:2509.20358*, 2025a.

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

Xi Wang, Robin Courant, Marc Christie, and Vicky Kalogeiton. Akira: Augmentation kit on rays for optical video generation. In *CVPR*, 2025c.

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

Jason Wei. Asymmetry of verification and verifier's rule. https://www.jasonwei.net/blog/asymmetry-of-verification-and-verifiers-law, 2025. Accessed: 2025-09-01.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. *arXiv preprint arXiv:2505.13934*, 2025.

Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative cartoon animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10793–10804, 2025.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. PhyT2V: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024a.

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

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. *arXiv preprint arXiv:2510.11512*, 2025.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley Chan. NewtonGen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. In *International Conference on Learning Representations*, pp. 12847–12878, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/15a321fbfc19fce9b325ec6e77dfb597-Paper-Conference.pdf.

Chenyu Zhang, Daniil Cherniavskii, Andrii Zadaianchuk, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025a.

David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024a.

Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: LLMs-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025b.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. PhysRVG: Physics-aware unified reinforcement learning for video generative models. In *Computer Vision – ECCV 2026*, pp. 149–166, Cham, 2026. Springer Nature Switzerland. doi: 10.1007/978-3-032-37252-9_9. URL https://doi.org/10.1007/978-3-032-37252-9_9.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025c.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024b.

Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis. *arXiv preprint arXiv:2503.20822*, 2025.

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

APPENDIX

This Appendix provides additional results and analyses supporting the main paper. Appendix A presents the *full related work discussion*, with detailed context on prior approaches to physics-aware video generation. Appendix B provides *implementation details* of `NewtonBench-60K` (simulation pipeline, ramp construction, mask extraction) and the *full equations* for the evaluation metrics. Appendix C provides the *full out-of-distribution (OOD) evaluation* on the OOD split of `NewtonBench-60K`, with per-metric results across all five metrics. Appendix E presents a *real-world evaluation* on 361 free-fall videos from the PISA (Li et al., 2025b) benchmark, showing that our physics-grounded post-training, developed entirely in simulation, transfers to natural scenes and real gravitational motion. In Appendix F, we evaluate the reliability of optical flow in our framework and show that *optical flow is a good proxy of the velocity*. Appendix G provides *additional quantitative evidence* for **Finding 2**, demonstrating physical properties conservation reward prevents reward hacking by avoiding degenerate zero-motion solutions. Appendix H offers *qualitative comparisons* on OpenSora and Wan, together with extended Wan results across all five NMPs, with frame-level visualizations highlighting the improved consistency and stability of `NewtonRewards`.

A RELATED WORKS

Video Generative Models. Video models are progressing rapidly (Singer et al., 2022; Ho et al., 2022b;a; Blattmann et al., 2023a; Girdhar et al., 2023; Bar-Tal et al., 2024; Brooks et al., 2024; Blattmann et al., 2023b; Wang et al., 2025d; Yu et al., 2023). Using large-scale datasets and scalable backbones (e.g., DiT-style architectures (Peebles & Xie, 2023)), modern models produce photorealistic short clips conditioned on text, images, and other control signals (Lee et al., 2022; Wang et al., 2025c; Ma et al., 2025; Peng et al., 2024; Zhang et al., 2024b;a; Guo et al., 2024; Bhowmik et al., 2025). Open-source efforts, such as OpenSora (Zheng et al., 2024), CogVideo (Hong et al., 2023; Yang et al., 2024b), HunyuanVideo (Kong et al., 2024), and Wan 2 (Wan et al., 2025), have shown big improvements in visual fidelity and conditional control. Although these models show some emergent reasoning on tasks beyond their training (Wiedemer et al., 2025), scaling data or model size alone often cannot eliminate unrealistic motion or physics violations (Kang et al., 2025; Motamed et al., 2025b).

Physics-aware Video Generation. There are 3 groups of strategies for physics priors/dynamics in video generation.

(1) Instruction and feedback-based fine-tuning. Several methods fine-tune video diffusion models using feedback or textual instructions from LLMs (Lian et al., 2023; Zhang et al., 2025b; Hao et al., 2025) or VLMs (Yang et al., 2025; Motamed et al., 2025a). Others align generator representations with self-supervised video encoders (Zhang et al., 2025c), whose features exhibit intuitive physical understanding (Garrido et al., 2025). For instance, PhyT2V (Xue et al., 2025) introduces a feedback loop where an LLM checks if generated videos obey physical laws, reasoning over captions from a video captioning model. However, such feedback is indirect and often reflects perception rather than physics consistency.

(2) Physics-guided simulation and representation. Another line of work leverages physics simulators or 3D representations to guide video generation. Some approaches pre-compute 3D or physically plausible conditions as auxiliary input for the video generator (Xie et al., 2025; Wang et al., 2025a; Yuan et al., 2026). Others build integrated pipelines combining simulation, rendering, and generative modeling to ensure motion realism (Lv et al., 2024; Liu et al., 2024; Hsu et al., 2024). For instance, PhysGen (Liu et al., 2024) performs image-based warping with simulated motion dynamics given user-defined forces and torques. However, relying on external simulators enforces physics only indirectly, through the pre-trained model's interpretation of conditioned data. *(3) Physics-rich datasets and post-training.* This strategy focuses on data-level physical grounding. PISA (Li et al., 2025b) introduced a dataset of 361 real-world and 60 synthetic videos of objects falling in diverse environments and used post-training strategies such as supervised fine-tuning and object-reward optimization, aligning optical flow, depth, and segmentation maps to improve physical consistency. Methods such as (Zhao et al., 2025; Wang et al., 2025b; Chen et al., 2025) rely on physics-rich or instruction-grounded datasets to enhance physical plausibility in a data-driven manner. Similarly,

PhysMaster (Ji et al., 2025) proposed optimizing a neural *PhysEncoder* via reinforcement learning, using costly human-annotated preference data.

Recent Advances and Comparison Scope. Recent works further explore physical dynamics in video generation through distinct mechanisms. NewtonGen (Yuan et al., 2026) learns Neural Newtonian Dynamics and converts predicted physical states into optical-flow controls at inference; physics enters through motion conditioning, rather than reward-based post-training of the video generator. PHANTOM (Shen et al., 2026) augments Wan with a parallel physics branch that jointly predicts V-JEPA 2 representations and video latents through flow matching, learning latent dynamics without an explicit equation-of-motion penalty. PhysRVG (Zhang et al., 2026) also uses reward-based post-training, but optimizes agreement with reference object trajectories, with collision-aware weighting, through a Mimicry-Discovery Cycle. In contrast, `NewtonRewards` directly penalizes a constant-acceleration residual computed from optical-flow proxies of generated videos, complemented by a physical properties conservation reward. Our kinematic term measures adherence to a governing constraint rather than agreement with a particular reference trajectory, while the properties term uses reference features to preserve object consistency. These rewards optimize the generator without adding a dynamics branch or requiring motion-control inputs at inference. Our comparison in Table 2 fixes the SFT initialization, generator architecture, training data, and conditioning within each backbone to isolate the effect of reward formulation on the same five NMPs.

B NewtonBench-60K: Implementation Details

B.1 Simulation Pipeline

We build upon *Kubric* (Greff et al., 2022) for scene orchestration, *PyBullet* for rigid-body dynamics, and *Blender* for rendering. Gravity is $\mathbf{g} = (0, 0, -9.81)$; videos are rendered at 512×512 resolution, 32 frames at 16 fps with HDRI lighting. We adopt a fixed side-view camera for unambiguous 2D motion analysis; per-clip outputs include RGB frames, instance masks, depth maps, and metadata (camera intrinsics/extrinsics and object attributes).

B.2 Assets and Splits

Objects are sampled from the GSO dataset (Downs et al., 2022) and backgrounds from HDRI (Greff et al., 2022). We create disjoint train/test pools for both objects and backgrounds, sampling strictly within the selected split.

B.3 Ramp Construction

Ramps are built from rectangular colliders (for stable physics) and visually aligned slabs (for rendering), with the top plane and downhill direction computed from the ramp mesh to ensure consistent tangential placement and motion direction.

B.4 Ramp Sliding: Detailed Force Decomposition

For a ramp inclined by angle θ relative to the horizontal, let $\hat{\mathbf{s}} = (t_x, t_y)$ be the unit tangent vector along the ramp's downhill direction in the image plane, $\hat{\mathbf{n}}$ the in-plane normal, and v_s the object's velocity component along $\hat{\mathbf{s}}$. The downhill component of gravity is $mg\sin\theta\,\hat{\mathbf{s}}$, and kinetic friction opposes the motion, $\mathbf{F}_f = -\mu_k mg\cos\theta\,\mathrm{sgn}(v_s)\,\hat{\mathbf{s}}$. The net tangential force along $\hat{\mathbf{s}}$ is therefore

$$\mathbf{F}_s = mg\sin\theta - \mu_k mg\cos\theta\,\mathrm{sgn}(v_s),$$

and the tangential acceleration is

$$\mathbf{a}_s = g\left(\sin\theta - \mu_k\cos\theta\,\mathrm{sgn}(v_s)\right) = \begin{cases} g(\sin\theta - \mu_k\cos\theta), & \text{sliding down } (v_s > 0), \\ g(\sin\theta + \mu_k\cos\theta), & \text{sliding up } (v_s < 0). \end{cases}$$

When sliding up, gravity and friction both act downhill, so $\mathbf{a}_s > 0$ opposes the uphill velocity and decelerates the object. The mass m cancels, so the kinematics depend only on g, θ, μ_k, and the initial state. $\mathbf{a}_s$ is constant within each sliding phase. If an object sliding up comes to rest and slides back, $\mathbf{a}_s$ switches once to the sliding-down value. The projection onto the image plane is $\mathbf{a} = (\mathbf{a}_s t_x, \mathbf{a}_s t_y)$.

Table 3: Out-of-distribution (OOD) evaluation on the 5K OOD benchmark of `NewtonBench-60K`. `NewtonRewards` improves consistently across all metrics, demonstrating stronger generalization than the OpenSora (SFT) baseline. ID rows are reproduced from the main paper for direct comparison.

Method	L2 (↓)	CD (↓)	IoU (↑)	$\text{RMSE}_{\mathbf{v}}$ (↓)	$\text{RMSE}_{\mathbf{a}}$ (↓)	Avg. Change
OpenSora (SFT) – ID	0.1098	0.3159	0.1103	0.2792	3.3244	–
`NewtonRewards`– ID	**0.0962** (+12.39%)	**0.2930** (+7.25%)	**0.1266** (+14.78%)	**0.2628** (+5.87%)	**3.0432** (+8.46%)	**+9.75%**
OpenSora (SFT) – OOD	0.1297	0.4082	0.0998	0.4230	6.1451	–
`NewtonRewards`– OOD	**0.1206** (+7.02%)	**0.3780** (+7.40%)	**0.1025** (+2.71%)	**0.3816** (+9.79%)	**5.1561** (+16.09%)	**+8.60%**

B.5 Mask Extraction

For ground truth, we use renderer instance masks to obtain per-frame object regions. For generated videos, we extract object masks with SAM2 (Ravi et al., 2025) guided by conditioning frames and object prompts; centroids $\mathbf{c}^{\text{gen}}$ are then computed from these masks for metric evaluation. This procedure provides a consistent, annotation-free way to track the same physical object across frames for both simulated and generated clips.

B.6 Evaluation Metric Equations

All metrics are computed per object (frame-aligned), then averaged across objects and videos. For frame interval Δt, $\mathbf{c}_t^{\text{gen}}, \mathbf{c}_t^{\text{gt}} \in \mathbb{R}^2$ are generated and ground-truth centroids at frame t.

Velocity RMSE. Image-plane velocities are defined by the first discrete derivative $\mathbf{v}_t = \frac{\mathbf{c}_{t+1} - \mathbf{c}_t}{\Delta t}$, and the velocity error measures Newtonian consistency of first-order kinematics:

$$\text{RMSE}_{\mathbf{v}} = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T-1} \left\| \mathbf{v}_t^{\text{gen}} - \mathbf{v}_t^{\text{gt}} \right\|_2^2}.$$

Acceleration RMSE. Image-plane accelerations are defined by the discrete second-order derivative $\mathbf{a}_t = \frac{\mathbf{v}_{t+1} - \mathbf{v}_t}{\Delta t} = \frac{\mathbf{c}_{t+2} - 2\mathbf{c}_{t+1} + \mathbf{c}_t}{\Delta t^2}$, giving

$$\text{RMSE}_{\mathbf{a}} = \sqrt{\frac{1}{T-2} \sum_{t=1}^{T-2} \left\| \mathbf{a}_t^{\text{gen}} - \mathbf{a}_t^{\text{gt}} \right\|_2^2}.$$

These two physics metrics evaluate whether generated motions obey constant acceleration on both axes.

Trajectory Position Error. $\text{L2_traj} = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{c}_t^{\text{gen}} - \mathbf{c}_t^{\text{gt}}\|_2$.

CD and IoU. Chamfer Distance (CD) is computed between per-frame binary masks; Intersection over Union (IoU) is the per-frame mask overlap. Together with L2_traj, these capture pixel-space alignment and shape agreement.

C Out-of-Distribution Generalization

We evaluate generalization under distribution shift on the 5K OOD split of `NewtonBench-60K`, where test videos exhibit higher drop heights, faster throws, steeper ramps, and perturbed friction compared to training. The OpenSora (SFT) baseline degrades substantially in this setting (e.g., L2 increases to 0.1297 and acceleration error nearly doubles to 6.15), reflecting limited robustness to unseen physical configurations. In contrast, `NewtonRewards` consistently improves across all five metrics–achieving a $+7.02\%$ reduction in L2, $+7.40\%$ improvement in CD, and $+9.79\%$ reduction in velocity error–despite never observing OOD dynamics during post-training. These results show that enforcing Newtonian kinematic and physical properties constraints yields models that not only fit training physics more faithfully but also extrapolate more reliably to unseen physical regimes.

Table 4: Comparison of different post-training strategies on the OpenSora and Wan 2.2 5B SFT baselines. Percentages indicate relative change vs. the corresponding baseline (green = improvement, red = regression). Visual metrics (L2, CD, IoU) capture pixel alignment and shape agreement; physics metrics ($\text{RMSE}_\mathbf{v}$, $\text{RMSE}_\mathbf{a}$) capture physical plausibility in motion.

	Visual metrics			Physics metrics		
Method	**L2** (↓)	**CD** (↓)	**IoU** (↑)	$\textbf{RMSE}_\mathbf{v}$ (↓)	$\textbf{RMSE}_\mathbf{a}$ (↓)	**Avg. Change**
OpenSora (SFT)	0.1098	0.3159	0.1103	0.2792	3.3244	–
PISA (Li et al., 2025b) ORO Optical Flow	0.1042 (+5.10%)	0.2963 (+6.18%)	0.1179 (+6.88%)	0.2799 (-0.25%)	2.7217 (+18.12%)	+7.21%
PISA (Li et al., 2025b) ORO Depth Map	0.1079 (+1.73%)	0.3114 (+1.43%)	0.1146 (+3.90%)	0.2875 (-2.97%)	3.4652 (-4.23%)	-0.03%
PISA (Li et al., 2025b) ORO Segmentation	0.1099 (-0.09%)	0.3177 (-0.57%)	0.1138 (+3.17%)	0.2796 (-0.14%)	3.2943 (+0.91%)	+0.65%
`NewtonRewards`	**0.0962** (+12.39%)	**0.2930** (+7.25%)	**0.1266** (+14.78%)	**0.2628** (+5.87%)	3.0432 (+8.46%)	**+9.75%**
`NewtonRewards` (w/o kinematic)	0.1109 (-1.00%)	0.3199 (-1.27%)	0.1145 (+3.81%)	0.2793 (-0.04%)	3.3321 (-0.23%)	+0.25%
`NewtonRewards` (w/o physical properties)	0.1055 (+3.92%)	0.2993 (+5.26%)	0.1165 (+5.62%)	0.2737 (+1.97%)	**2.5348** (+23.75%)	+8.10%
Wan 2.2 5B (SFT)	0.0650	0.2468	0.3316	0.3767	8.7004	–
PISA (Li et al., 2025b) ORO Optical Flow	0.0655 (-0.77%)	0.2145 (+13.09%)	0.3430 (+3.44%)	0.3821 (-1.43%)	7.6637 (+11.92%)	+5.25%
PISA (Li et al., 2025b) ORO Depth Map	0.0739 (-13.69%)	0.2239 (+9.28%)	0.3172 (-4.34%)	0.3805 (-1.01%)	8.7249 (-0.28%)	-2.01%
PISA (Li et al., 2025b) ORO Segmentation	0.0639 (+1.69%)	**0.1932** (+21.72%)	0.3401 (+2.56%)	0.3001 (+20.33%)	8.7944 (-1.08%)	+9.05%
`NewtonRewards`	0.0625 (+3.85%)	0.2158 (+12.56%)	**0.3466** (+4.52%)	**0.2929** (+22.25%)	**6.5561** (+24.65%)	**+13.56%**
`NewtonRewards` (w/o kinematic)	0.0649 (+0.15%)	0.2546 (-3.16%)	0.3354 (+1.15%)	0.3701 (+1.75%)	8.2863 (+4.76%)	+0.93%
`NewtonRewards` (w/o physical properties)	**0.0620** (+4.62%)	0.2171 (+12.03%)	0.3460 (+4.34%)	0.3127 (+16.99%)	7.1495 (+17.83%)	+11.16%

D Comparison to State of the Art

We report in Table 4 the full comparison to the state of the art, including relative changes per category.

E Real-World Experiment

To test whether physics-grounded post-training transfers beyond controlled simulation, we evaluate our approach on 361 real-world free-fall videos provided in the PISA benchmark (Li et al., 2025b). These videos capture everyday objects dropped in diverse indoor and outdoor environments, providing natural variation in texture, lighting, background clutter, and real gravitational motion (see Figure 6). This setting lets us evaluate whether a model trained purely on synthetic, physics-informed data can generalize to real-world scenarios.



Figure 6: **Real-world free-fall evaluation.** We test whether models post-trained purely in simulation can generalize to real camera imagery and real gravitational motion. Here is a representative video from the PISA(Li et al., 2025b) real-world dataset (361 free-fall videos).

Following the evaluation protocol used in our `NewtonBench-60K`, we measure performance using five metrics: three visual metrics (L2, Chamfer Distance, IoU) and two physics metrics ($\text{RMSE}_\mathbf{v}$, $\text{RMSE}_\mathbf{a}$). We evaluate the OpenSora (SFT) baseline, all three PISA post-training variants (Depth, Segmentation, Optical Flow), and our `NewtonRewards` under exactly the same conditions. This establishes a direct comparison of how well different reward formulations cope with real free-fall trajectories.

As shown in Table 5, `NewtonRewards` yields the largest and maintains consistent gains across both visual and physics metrics, demonstrating strong transfer from synthetic supervision to natural free-fall motion.

F Reliability of Optical Flow

Optical flow is a good proxy of the velocity. We compare RAFT(Teed & Deng, 2020) optical flow on synthetic data with the ground-truth object velocity (from simulation). In 97% of videos, the

Table 5: SFT and post-training strategies on real-world experiment. (green = improvement, red = regression). Visual metrics (L2, CD, IoU) capture pixel alignment and shape agreement; physics metrics ($\mathrm{RMSE}_{\mathbf{v}}$, $\mathrm{RMSE}_{\mathbf{a}}$) capture physical plausibility in motion.

	Visual metrics			Physics metrics		
Method	**L2** (↓)	**CD** (↓)	**IoU** (↑)	$\mathbf{RMSE_v}$ (↓)	$\mathbf{RMSE_a}$ (↓)	**Avg. Change**
OpenSora (SFT)	0.1716	0.4386	0.0198	2.4485	18.4169	–
PISA (Li et al., 2025b) ORO Optical Flow	0.1699 (+0.99%)	0.4336 (+1.14%)	0.0182 (-8.08%)	2.4237 (+1.01%)	18.3333 (+0.45%)	-0.90%
PISA (Li et al., 2025b) ORO Depth Map	0.1704 (+0.70%)	0.4342 (+1.00%)	0.0218 (+10.10%)	2.4395 (+0.37%)	18.3474 (+0.38%)	+2.51%
PISA (Li et al., 2025b) ORO Segmentation	0.1712 (+0.23%)	0.4372 (+0.32%)	0.0210 (+6.06%)	2.4273 (+0.87%)	18.2344 (+0.99%)	+1.69%
`NewtonRewards`	**0.1698** (+1.05%)	**0.4333** (+1.21%)	**0.0235** (+18.69%)	**2.3889** (+2.43%)	**18.1670** (+1.36%)	**+4.95%**

ground-truth velocity lies within the 95% confidence interval of the estimation, validating that optical flow is a sufficient and good proxy for the kinematic reward.

We further evaluate optical flow reliability by comparing the flow on generated videos with that from the simulation. Table 6 shows that `NewtonRewards` achieves the lowest absolute and relative flow error (0.0855 / 8.43%), outperforming SFT baseline and PISA (Li et al., 2025b) Optical Flow variant. This gain comes from the physical properties conservation reward, which stabilizes object physical properties by discouraging object deformation (physical properties consistency). Thus, our designed rewards improve the reliability of optical flow as a velocity proxy, partially mitigating the imperfect video quality.

Table 6: **Optical flow error on generated videos.** Lower is better.

	OpenSora (SFT)	**PISA (Optical Flow)**	`NewtonRewards`
Optical Flow Error ↓	0.1177 (14.25%)	0.1034 (12.98%)	**0.0855 (8.43%)**

Our method is formulated in an Eulerian framework, using optical flow at fixed pixel locations, rather than explicitly tracking Lagrangian material points. In our simulation setting, the camera is fixed, the background is static, and object motion occurs in a plane parallel to the image plane. Under these assumptions, Eulerian optical flow has been widely used in the computer vision community as a reliable proxy for object velocity.

G Reward Hacking Mitigation

Table 7: **Residuals and velocity magnitude.** Lower residuals indicate smoother (closer-to-constant-acceleration) motion, while the velocity magnitude reveals whether the motion remains physically meaningful. Without physical properties conservation, the model reduces the residual primarily by collapsing motion magnitude.

Method	**Horizontal Residual** ↓	**Vertical Residual** ↓	**Velocity Magnitude**
OpenSora (SFT)	0.000509	0.001142	0.101715
`NewtonRewards`	**0.000194**	0.000511	0.071377
`NewtonRewards` w/o physical properties	0.000986	**0.000331**	**0.033854**

In the Experiments section of the main paper, we identified *reward hacking* as a failure mode that arises when optimizing only the kinematic residual. Without the physical properties-conservation reward, the generator can trivially reduce $\|\boldsymbol{\phi}_{t+1} - 2\,\boldsymbol{\phi}_t + \boldsymbol{\phi}_{t-1}\|_2^2$ by driving all velocity fields $\boldsymbol{\phi}_t$ toward zero, producing videos in which the object barely moves or even disappears. Here, we provide additional quantitative evidence for this finding.

As shown in Table 7, removing the physical properties conservation reward (`NewtonRewards` w/o physical properties) decreases the vertical residual compared to SFT, but does so by collapsing the average velocity magnitude from 0.1017 to 0.0339, a reduction of more than 66%. This confirms that the residual-only model optimizes the objective by freezing motion rather than by producing more accurate Newtonian trajectories. In contrast, the full `NewtonRewards` not only yields substantially lower residuals in both directions, but also maintains non-trivial velocity, indicating that it improves dynamical consistency without sacrificing meaningful motion. These results quantitatively support

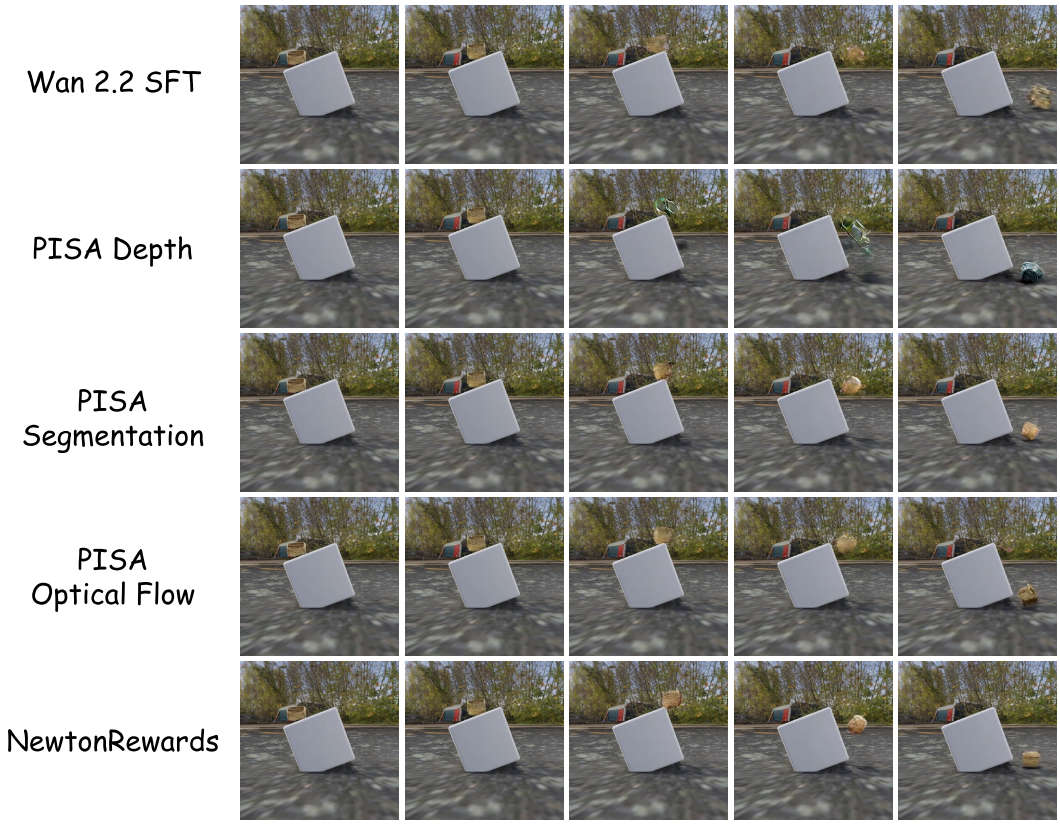

Figure 7: **Qualitative comparison of post-training strategies on Wan 2.2 5B in a `NewtonBench-60K` ramp scene.** Frames progress from left to right. SFT and PISA variants exhibit changes in the moving object's appearance or irregular displacement across the sequence. In contrast, `NewtonRewards` better preserves the object's appearance and produces a more coherent trajectory, complementing the quantitative results in Table 2.

our claim that the physical properties conservation reward is crucial for preventing reward hacking and stabilizing physics-grounded post-training.

H Extended Qualitative Results

Figures 7 and 8 provide representative comparisons on Wan 2.2 5B and OpenSora, respectively. In Figure 8, objects either drift unnaturally or exhibit inconsistent acceleration when interacting with the ramp. For example, PISA Depth (Row 2) shows the cube briefly losing contact in Frame 3, and PISA OF (Row 4) produces a sudden orientation snap between Frames 3–4. In contrast, `NewtonRewards` maintains stable contact and smoother motion. The Wan comparison in Figure 7 similarly highlights differences in object appearance and trajectory coherence.

In addition, we compare post-training strategies on Wan 2.2 5B across all five NMPs in figs. 9 to 13. Columns show SFT, PISA (Li et al., 2025b) Depth, PISA Segmentation, PISA Optical Flow, and `NewtonRewards`, respectively. Each comparison uses the same scene and conditioning prefix from `NewtonBench-60K`.

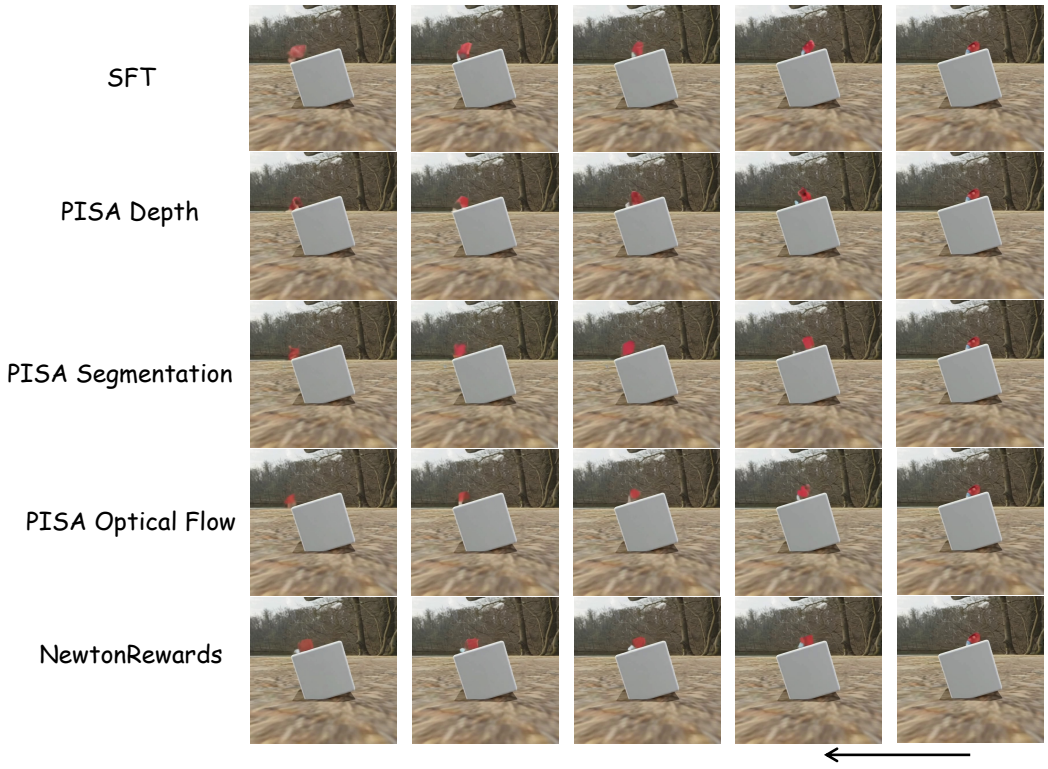

Figure 8: **Qualitative comparison of post-training strategies on OpenSora in the `NewtonBench-60K` ramp-slide down scenario.** Clear differences emerge when inspecting the temporal evolution across frames (right→left). For SFT and all PISA variants (Depth, Seg, Optical Flow (OF)), the cube exhibits inconsistent deceleration and unstable surface contact-evident in Frames 2–4, where the cube tilts unnaturally, slips erratically, or momentarily "floats" above the ramp. PISA OF especially shows noticeable jitter and non-smooth frame-to-frame motion. In contrast, `NewtonRewards` maintains stable grounding and smooth, constant-acceleration motion across all frames.



Figure 9: **Qualitative comparison of post-training strategies on Wan 2.2 5B in the `NewtonBench-60K` free fall scenario (F).** Rows show frames from top to bottom. Differences emerge during descent and ground contact: SFT blurs several objects, while PISA Depth and Segmentation introduce noticeable changes in object shape and color. In contrast, `NewtonRewards` better preserves recognizable objects across the displayed continuation.



Figure 10: **Qualitative comparison of post-training strategies on Wan 2.2 5B in the `NewtonBench-60K` horizontal throw scenario (TH).** Rows show frames from top to bottom. PISA Depth and Segmentation visibly alter the objects' appearance after conditioning, while SFT makes them less distinct. In contrast, `NewtonRewards` retains both objects and provides a more coherent continuation of their rightward and downward motion.



Figure 11: **Qualitative comparison of post-training strategies on Wan 2.2 5B in the `NewtonBench-60K` parabolic throw scenario (TP).** Rows show frames from top to bottom. PISA Depth introduces spurious fragments and substantial appearance changes, while the blue object becomes less distinct under SFT. In contrast, `NewtonRewards` better preserves recognizable objects and their spatial progression across the displayed continuation.



Figure 12: **Qualitative comparison of post-training strategies on Wan 2.2 5B in the `NewtonBench-60K` ramp slide down scenario (RD).** Rows show frames from top to bottom. As the moving objects emerge from behind the ramp, PISA Depth introduces conspicuous changes in color and shape. In contrast, `NewtonRewards` maintains a more consistent object appearance during the visible descent, while preserving the ramp and background.



Figure 13: **Qualitative comparison of post-training strategies on Wan 2.2 5B in the `NewtonBench-60K` ramp slide up scenario (RU).** Rows show frames from top to bottom. PISA Depth produces pronounced shape distortion, and PISA Optical Flow stretches the moving object in the final displayed frame. In contrast, `NewtonRewards` better preserves its compact structure as it moves along and beyond the upper edge of the ramp.