What about gravity in video generation? Post-Training Newton's Laws with Verifiable Rewards

Minh-Quan Le[1] Yuanzhi Zhu[2] Vicky Kalogeiton[2,†] Dimitris Samaras[1,†]
[1]Stony Brook University [2]LIX, École Polytechnique, CNRS, IPP [†]Equal Advising
https://cvlab-stonybrook.github.io/NewtonRewards

Abstract

Recent video diffusion models can synthesize visually compelling clips, yet often violate basic physical laws–objects float, accelerations drift, and collisions behave inconsistently–revealing a persistent gap between visual realism 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 appearance features serve as a proxy for mass. These proxies enable explicit enforcement of Newtonian structure through two complementary rewards: a Newtonian kinematic constraint enforcing constant-acceleration dynamics, and a mass 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 our newly constructed large-scale benchmark,* `NewtonBench-60K`*. Across all primitives 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 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 [3, 30, 31, 38, 43, 71]. Recent generative models can produce visually stunning videos from text [15, 21, 52, 66, 77], images [7, 15, 21, 52], or latent trajectories [12], but often exist in worlds unbound by physics: worlds where objects float [69], collisions resolve unrealistically [3], and motion unfolds without cause [71]. These violations of basic laws, such as Newton's laws of motion, highlight a lack of *physical realism* in video generation.

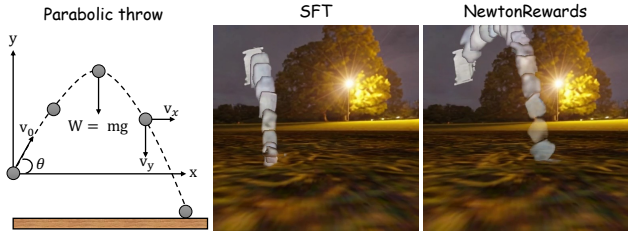

Figure 1. `NewtonRewards` enforce physical laws in video generation. Shown is a parabolic throw scenario from our NewtonBench-60K dataset. Baseline supervised fine-tuning (SFT) produces implausible motion violating Newtonian dynamics. Our `NewtonRewards` post-training restores parabolic trajectories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

Embedding physical plausibility into video generation is more than an aesthetic choice; it is a necessity. In many applications, from immersive game environments and realistic cinematic worlds [55], to training world models for games [10, 29, 67], autonomous driving [1, 14, 27, 48] and robotic control [4, 35, 64], generated videos serve as data for perception, reasoning, and action [60]. In these scenarios, non-physical dynamics can lead to inconsistent learning signals, unrealistic affordances, and failure to generalize to the real world. 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.

Recently, several approaches have sought to embed physical plausibility into video generation. These range from methods that fine-tune diffusion models using textual or feedback-based supervision from large language or vision-language models [16, 22, 33, 42, 54, 62, 65, 73, 74], to approaches that incorporate physical simulators or 3D representations as motion or geometry priors [26, 36, 39, 40, 53, 61, 70, 76], and to those that rely on physics-rich datasets or post-training signals derived from real or syn-

thetic videos [11, 28, 36, 54]. Still, these methods typically use physics signals coming from humans or VLM feedback as a condition rather than explicitly enforcing physics laws.

Neither humans nor "VLMs-as-judge" can precisely evaluate how well physical constraints are being followed (apart from egregious physical law violations). As a result, generated videos often appear visually realistic yet fail to satisfy physics principles such as momentum conservation, force–acceleration proportionality, and consistent gravitational effects. We argue that despite improved perceptual realism and motion smoothness in recent work, physically plausible video generation remains a challenge, as models must respect the physical laws governing object dynamics.

Therefore, in this work, we propose the first verifiable rewards for physical laws in video generation, which employ rule-based evaluations that can automatically verify the correctness of the output [19, 32, 44, 49, 57, 58]. Given a video generator, physical quantities such as velocity or force cannot be directly observed from its raw output frames. To bridge this gap, we estimate these quantities using pre-trained utility models (e.g., optical flow or video-embedding networks). Their outputs, which we term *measurable proxies*, serve as observable surrogates for underlying physical variables. By defining physics constraints and rewards on these proxies, we post-train video generators to produce videos that follow physical laws explicitly.

As an initial application of this approach, we present `NewtonRewards`, as a basis for a framework for Newtonian motion constraints. We define measurable proxies (optical flow and appearance embeddings) for physical constraints (velocity and mass) and use them to formulate both kinematic and mass-conservation rewards. We evaluate our framework across five Newtonian Motion Primitives (NMPs): (i-iii) *free fall, horizontal and parabolic throw*, and (iv-v) *ramp sliding down/up*. We experiment on our large-scale simulated dataset, **NewtonBench-60K**, specifically designed for dynamic motion evaluation, with diverse scenarios for each NMP. Empirically, `NewtonRewards` consistently improves physical plausibility, motion smoothness, and temporal coherence over prior methods such as PISA [36]. It yields gains across all five NMPs for both in-distribution (ID) and out-of-distribution (OOD) settings. Physics-grounded constraints correct kinematic violations that visual feature alignment alone cannot fix, i.e., reducing constant-acceleration residuals and mitigating reward hacking behaviors (objects vanish to minimize motion).

Our contributions can be summarized as:

- We introduce `NewtonRewards`, an elegant physics-grounded post-training framework for video generation that explicitly enforces Newtonian dynamic motions (e.g., throw, free fall, slide with friction).
- We employ optical flow and visual features as a differentiable proxy to devise Newtonian kinematic and mass constraints, yielding verifiable, rule-based rewards that promote physically correct motion.
- 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.
- Experiments show `NewtonRewards` consistently improves across both visual and physics metrics, across all five NMPs, for both ID and OOD settings, producing more physically faithful and temporally coherent videos.

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.

2. Related Works

Video Generative Models. Video models are progressing rapidly [5, 7–9, 17, 23, 24, 50, 56, 68]. Using large-scale datasets and scalable backbones (e.g., DiT-style architectures [45]), modern models produce photorealistic short clips conditioned on text, images, and other control signals [6, 20, 34, 41, 46, 55, 72, 75]. Open-source efforts, such as OpenSora [77], CogVideo [25, 66], HunyuanVideo [31], and Wan 2.1 [52], have shown big improvements in visual fidelity and conditional control. Although these models show some emergent reasoning on tasks beyond their training [59], scaling data or model size alone often cannot eliminate unrealistic motion or physics violations [30, 43].

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 Large Language Models (LLMs) [22, 37, 73] or Vision-Language Models (VLMs) [16, 42, 65, 74]. For instance, PhyT2V [62] 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 precompute 3D or physically plausible conditions as auxiliary input for the video generator [53, 61, 70]. Others build integrated pipelines combining simulation, rendering, and generative modeling to ensure motion realism [26, 39, 40, 62]. For instance, PhysGen [39] 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.

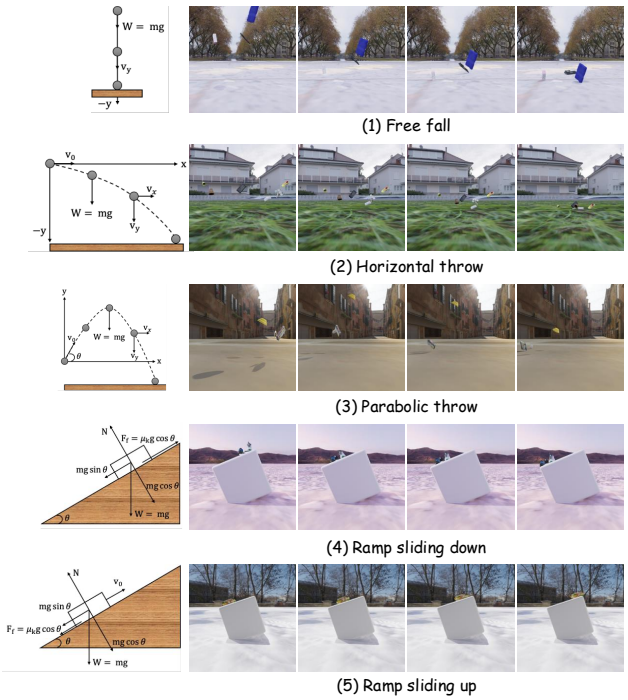

Figure 2. Illustration of the five NMPs in the proposed `NewtonBench-60K` dataset. Left: corresponding free-body diagrams showing dominant forces and accelerations. Right: rendered trajectories from our Kubric-based [18] simulator, demonstrating constant-acceleration dynamics in diverse environments.

(3) Physics-rich datasets and post-training. This strategy focuses on data-level physical grounding. PISA [36] 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 [11, 54, 76] rely on physics-rich or instruction-grounded datasets to enhance physical plausibility in a data-driven manner. Similarly, PhysMaster [28] proposed optimizing a neural *PhysEncoder* via reinforcement learning, using costly human-annotated preference data.

3. Newtonian Motion Primitives (NMPs)

We formulate post-training of video diffusion models through classical mechanics, grounded in Newton's three laws of motion (Section 3.1). In our setting, an object observed in a video sequence undergoes motion determined by external forces acting upon it. Building upon these physical laws, 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, as described in detail in Section 3.2.

3.1. 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 First Law (Law of Inertia) states that an object remains at rest or continues in uniform motion unless acted upon by an external force. It underlies scenarios such as *free fall*, *horizontal throw*, and *parabolic throw*, where the absence of horizontal forces yields $\mathbf{a}_x = 0$ and constant $\mathbf{v}_x$.

Newton's Second Law (Law of Acceleration) relates the net force $\mathbf{F}_{\text{net}}$ to the resulting acceleration $\mathbf{a}$ and mass m:

$$\mathbf{a} = \mathbf{F}_{\text{net}}/m \quad . \tag{1}$$

It provides the quantitative foundation for our kinematic rewards, linking motion to underlying forces and mass.

Newton's Third Law (Action-Reaction) states that when two bodies interact, they exert equal and opposite forces on each other. In our context, when an object interacts with a ramp or surface, the ramp exerts an equal and opposite normal force that balances the contact, together with a tangential frictional force opposing motion.

3.2. 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 $x = (f/Z)X$ with proportional factor f/Z. For short intervals, this scale is nearly constant, allowing analysis in image-space. *

(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:

$$\mathbf{a}_x = 0 \quad , \qquad \mathbf{a}_y = -g \quad , \tag{2}$$

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).

(4-5) NMP-RD and NMP-RU: Ramp Sliding with Friction. For a ramp inclined by angle θ relative to the horizontal, $\hat{\mathbf{s}} = (t_x, t_y)$ is the unit tangent vector along the ramp's downhill direction in the image plane, and $\hat{\mathbf{n}}$ the in-plane normal. Decomposing the gravitational force with kinetic friction $\mathbf{F}_f = -\mu_k mg \cos\theta\, \hat{\mathbf{s}}$, the net tangential force is

$$\mathbf{F}_s = \pm\left(mg \sin\theta - \mu_k mg \cos\theta\right) \quad , \tag{3}$$

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

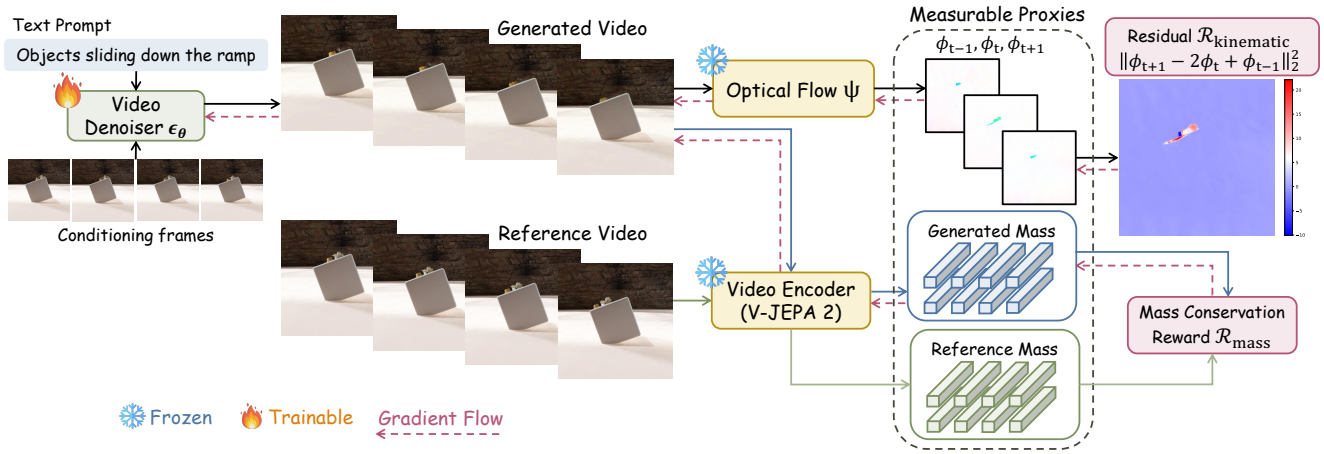

Figure 3. **Physics-Grounded Video Post-Training Pipeline.** Our method improves a pre-trained video generator by using physics-based rewards. Utility models (optical flow Ψ and V-JEPA 2) process the generated video to compute measurable proxies, from which kinematic and mass conservation rewards are derived to enforce explicit physics constraints.

where the positive/negative sign corresponds to sliding down/up. The constant tangential acceleration is

$$\mathbf{a}_s = \pm\, g(\sin\theta - \mu_k \cos\theta) \quad , \tag{4}$$

and its projection onto the image plane is $\mathbf{a} = (\mathbf{a}_s t_x, \mathbf{a}_s t_y)$.

4. 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 3): (i) computing measurable proxies from model outputs (Section 4.1), and (ii) defining reward functions that quantify adherence to physics constraints; Section 4.2 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.

4.1. Measurable Proxies

Let $G_\theta(\epsilon, c)$ denote a video generator parameterized by 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 4.1.1) and visual appearance embeddings as proxies for mass-related properties (Section 4.2.1). These proxies allow us to formulate physics-grounded constraints, such as constant-acceleration residuals and mass-conservation consistency, that provide verifiable reward signals for fine-tuning the generator.

4.1.1. 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 using video frames V_i:

$$\boldsymbol{\phi}_t = \mathbf{\Psi}(V_t, V_{t+1}) \quad , \tag{5}$$

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

$$\mathbf{v}_t \approx \boldsymbol{\phi}_t / \Delta t \quad , \tag{6}$$

with frame interval $\Delta t = 1/\text{FPS}$.

4.1.2. Visual Features as Mass Proxy

Newton's Second Law states that acceleration is inversely proportional to mass under fixed external force (Eq. 1), implying that heavier objects exhibit smoother, slower changes in motion. Though absolute mass is not directly observable in videos, appearance and texture cues often correlate with object identity, material, and thus effective mass. We define a *mass proxy* based on high-level visual representations extracted from a pre-trained video encoder. Let $\mathbf{z}_t = M_{\text{mass}}(V_t)$ denote the per-frame feature embedding obtained from the encoder, capturing consistent object-level appearance information over time. This embedding space provides a differentiable, semantically aligned measure of visual mass consistency that can be compared across time or between simulated and generated videos.

4.2. Physics Constraints and Rewards

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

$$\mathcal{L}_{\text{phys}} = \sum_{j=1}^{J} \lambda_j \cdot \ell(C_j[M_i(V)]) \quad , \tag{7}$$

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

4.2.1. 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

$$\mathbf{v}_{t+1} - 2\mathbf{v}_t + \mathbf{v}_{t-1} = \mathbf{0} \quad . \tag{8}$$

This discrete second-derivative constraint enforces the constant acceleration implied by Newton's Second Law.

Substituting the proxy in Equation 6 into Equation 8, we obtain the unified residual for all NMPs (see Section 3.2):

$$\boldsymbol{\phi}_{t+1} - 2\,\boldsymbol{\phi}_t + \boldsymbol{\phi}_{t-1} \approx \mathbf{0} \quad . \tag{9}$$

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:*

$$\mathcal{R}_{\textit{kinematic}} = \left\|\boldsymbol{\phi}_{t+1} - 2\,\boldsymbol{\phi}_t + \boldsymbol{\phi}_{t-1}\right\|_2^2 \approx \mathbf{0} \quad . \tag{10}$$

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

4.2.2. Mass 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, feature-level similarity constraint encourages mass preservation, so that the generator maintains consistent object appearance–hence consistent inferred mass–throughout the sequence and across domains. The mass conservation residual

$$\mathcal{R}_{\text{mass}} = \frac{1}{T}\sum_{t=0}^{T-1} \left\|\mathbf{z}_t^{\text{gen}} - \mathbf{z}_t^{\text{sim}}\right\|_2^2 \quad , \tag{11}$$

penalizes deviations between generated and simulated visual features. Minimizing $\mathcal{R}_{\text{mass}}$ encourages the generator to produce motions and appearances that obey mass-dependent dynamics implied by Newton's laws, complementing the kinematic constraint in Equation 10.

4.2.3. Post-Training Objective

Our final objective combines a Newtonian kinematic constraint (Prop. 1) and a mass-matching term (Eq. 11):

$$\mathcal{L}_{\text{phys}} = \lambda_{\text{kinematic}}\,\mathcal{R}_{\text{kinematic}} + \lambda_{\text{mass}}\,\mathcal{R}_{\text{mass}} \quad . \tag{12}$$

As the background is static and the camera is fixed in the construction of our dataset, the optical flow field $\boldsymbol{\phi}$ directly captures object motion, letting us compute the loss over the entire frame. This objective enforces Newtonian consistency without requiring explicit acceleration/depth supervision, unifying all five motion primitives (Section 3.2) under a single principle: under constant external forces, image-plane accelerations remain constant.

5. 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 Figure 2. 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 [36] focuses solely on free-fall and does not capture a broader range of Newtonian dynamics.

5.1. Simulation Pipeline and Canonical Setups

Physics, Rendering, and Outputs. We build upon *Kubric* 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).

Assets and Splits. Objects are sampled from the GSO dataset [13] and backgrounds from HDRI [18]. We create disjoint train/test pools for both objects and backgrounds, sampling strictly within the selected split.

Table 1. Natural motion primitive parameterization.

NMP	Description
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 $\theta \in [15°, 75°]$.
Ramp slide down	Ramp angle $\theta \in [15°, 45°]$, kinetic friction coefficients $(\mu_{\text{ramp}}, \mu_{\text{obj}}) \approx 0.06$; objects positioned near crest and released from rest.
Ramp slide up	Same ramp construction; initial uphill $v_0 \in [3, 4]$ m/s imparted along the tangent.

NMP Parameterization. We generate each primitive in Table 1, by sampling a small set of physically meaningful parameters; ramps are built from rectangular colliders (stable physics) and visually aligned slabs (rendering), with top plane and downhill direction computed from the ramp mesh for consistent tangential placement and motion direction.

Mask Extraction for Evaluation. For ground truth, we use renderer instance masks to obtain per-frame object regions. For generated videos, we extract object masks with SAM2 [47] guided by conditioning frames and ob-

ject prompts; centroids $\mathbf{c}^{\text{gen}}$ are then computed from these masks for metrics evaluation (Sec. 5.3).

5.2. 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: *higher* horizontal throws (e.g., $[1.7, 2.0]$ m/s), *higher* 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 $\pm 25\%$ in OOD to decouple appearance from dynamics.

5.3. Evaluation Metrics

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. Evaluation includes physics-based metrics (velocity and acceleration RMSE) and standard visual metrics (L2, CD, and IoU) on both in and out-of distribution splits.

Velocity RMSE. We define image-plane velocities by the first discrete derivative $\mathbf{v}_t = \frac{\mathbf{c}_{t+1} - \mathbf{c}_t}{\Delta t}$. Our velocity error measures Newtonian consistency of first-order kinematics: $\textbf{RMSE}_{\mathbf{v}} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1} \left\|\mathbf{v}_t^{\text{gen}} - \mathbf{v}_t^{\text{gt}}\right\|_2^2}$.

Acceleration RMSE. Likewise, we define image-plane accelerations by the discrete second-order derivative,

$$\mathbf{a}_t = \frac{\mathbf{v}_{t+1} - \mathbf{v}_t}{\Delta t} = \frac{\mathbf{c}_{t+2} - 2\mathbf{c}_{t+1} + \mathbf{c}_t}{\Delta t^2} \quad ,$$

we report $\textbf{RMSE}_{\mathbf{a}} = \sqrt{\frac{1}{T-2}\sum_{t=1}^{T-2} \left\|\mathbf{a}_t^{\text{gen}} - \mathbf{a}_t^{\text{gt}}\right\|_2^2}$. These two physics metrics directly evaluate if generated motions obey constant-acceleration behavior in both axes.

Standard Visual Metrics (following PISA). Spatial fidelity metrics *Trajectory Position Error (centroid L2)* $\text{L2\_traj} = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{c}_t^{\text{gen}} - \mathbf{c}_t^{\text{gt}}\|_2$, *Chamfer Distance* (CD) between binary masks (per frame), and *Intersection over Union* (IoU) (per-frame overlap), capture pixel-space alignment and shape agreement.

6. Experiments

6.1. Experimental Settings

We fine-tune our base text-to-video diffusion model OpenSora v1.2 [77] on our `NewtonBench-60K` dataset comprising 50K simulated videos across five NMPs, and fine-tune it to accept both text and the first 4 frames of a video as conditions. 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. We employ the RAFT [51] optical-flow model to compute motion fields and the V-JEPA 2 [2] encoder to extract visual features for mass alignment. Evaluation follows the `NewtonBench-60K` protocol.

6.2. Experimental Results

Comparison with State of The Art. We compare our `NewtonRewards` framework with three post-training strategies adapted from PISA [36]: *Optical Flow Reward*, *Depth Reward*, and *Segmentation Reward*. All methods are fine-tuned from the same OpenSora (SFT) baseline under identical settings on our `NewtonBench-60K` dataset. In PISA, each reward measures the similarity between the generated video and its simulated ground truth: RAFT [51] computes optical flow fields and minimizes their discrepancy, Depth-Anything-V2 [63] aligns predicted and true depth maps, and SAM2 [47] provides object masks for IoU-based supervision. Table 2 reports the results.

We observe that visual feature-based rewards improve appearance metrics, *i.e.*, Depth and Optical flow rewards slightly enhance L2, CD, and IoU; our Mass 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.*

As shown in Table 2, PISA ORO–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 mass conservation reward shows a similar trend (row `NewtonRewards` w/o residual). Segmentation and optical flow rewards slightly worsen velocity measures, suggesting that frame-level feature alignment alone cannot capture Newtonian dynamics. In contrast, `NewtonRewards` achieves consistent improvements across all five metrics (average $+9.75\%$), demonstrating that enforcing self-consistent kinematic and mass constraints provides a stronger inductive bias for physically grounded and temporally coherent video generation.

Evaluation Across Newtonian Motion Primitives. Fig. 4 shows the relative performance gain of each post-training strategy over the OpenSora (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 motions such as 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

Table 2. Comparison of different post-training strategies on the OpenSora (SFT) baseline. Percentages indicate relative change vs. 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 [36] ORO Optical Flow	0.1042 (+5.10%)	0.2963 (+6.18%)	0.1179 (+6.88%)	0.2799 (-0.25%)	2.7217 (+18.12%)	+7.61%
PISA [36] ORO Depth Map	0.1079 (+1.73%)	0.3114 (+1.43%)	0.1146 (+3.90%)	0.2875 (-2.97%)	3.4652 (-4.23%)	+0.37%
PISA [36] 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 residual)	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 mass)	0.1055 (+3.92%)	0.2993 (+5.26%)	0.1165 (+5.62%)	0.2737 (+1.97%)	**2.5348** (+23.75%)	+8.10%

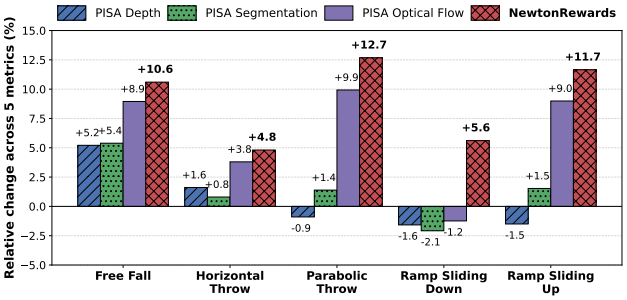

Figure 4. **Relative performance change across Newtonian Motion Primitives.** Percentage improvements over the SFT baseline across all five NMPs. Depth and Segmentation provide modest gains on simple motions but degrade on ramp dynamics, while Optical Flow shows highly variable and unstable behavior. In contrast, `NewtonRewards` delivers consistent positive improvements across all primitives, demonstrating robust generalization to diverse Newtonian dynamics.

regression on ramp-down despite large gains on free fall and parabolic throw. In contrast, `NewtonRewards` provides *uniformly positive* improvements across all five motion primitives. Its largest gains are on the most challenging motions–up to **+12.7%** on parabolic throws and **+11.7%** on ramp-up–and still outperforms all baselines on simple trajectories. Enforcing Newtonian kinematics and mass consistency yields robust, cross-regime improvements that generalize reliably across diverse physical scenarios.

Qualitative Comparison. Representative results across different post-training strategies are in Fig. 5. While PISA-based rewards grounded in visual similarity (Depth, Segmentation, Optical Flow) sometimes improve local appearance, they fail to enforce physically coherent motion: 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, our `NewtonRewards` yields visually realistic and physically consistent trajectories–objects maintain stable contact, decelerate smoothly under friction, and adhere to Newtonian expectations. Visual observations and quantitative trends in Table 2 and Fig. 6, confirm that physics-grounded verifiable

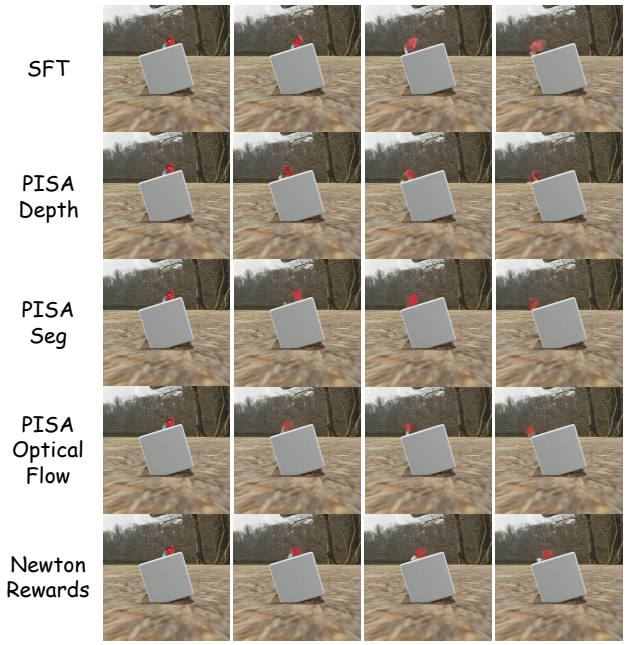

Figure 5. **Qualitative comparison of post-training strategies on the `NewtonBench-60K` ramp-slide down scenario.** Clear differences emerge when inspecting the temporal evolution across frames (left→right). For SFT and all PISA variants (Depth, Seg, Optical Flow), the cube exhibits inconsistent deceleration and unstable surface contact-evident in Frames 2–4, where the cube tilts unnaturally, slips erratically, or momentarily "floats" above the ramp. PISA Optical Flow 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.

rewards promote perceptual fidelity and dynamic realism.

OOD Evaluation. Table 3 evaluates generalization under distribution shift, 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 +7.01% reduction in L2, +7.38% improvement in CD, and +9.79% reduction

Table 3. Out-of-distribution (OOD) evaluation on 5K OOD benchmark of `NewtonBench-60K`. `NewtonRewards` improves consistently across all metrics, demonstrating stronger generalization than the OpenSora (SFT) baseline.

Method	L2 (↓)	CD (↓)	IoU (↑)	$\mathbf{RMSE_v}$ (↓)	$\mathbf{RMSE_a}$ (↓)	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%**

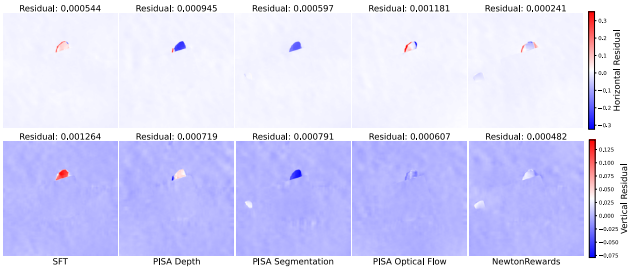

Figure 6. Mean horizontal and vertical residuals $\phi_{t+1} - 2\,\phi_t + \phi_{t-1}$. Lower magnitude indicates closer adherence to constant-acceleration dynamics. `NewtonRewards` produces the smallest residuals, while SFT and PISA variants show larger deviations.

in acceleration error; these, despite never observing OOD dynamics during post-training. These clearly show that enforcing Newtonian kinematic and mass constraints yields models that not only fit training physics more faithfully but also extrapolate more reliably to unseen physical regimes.

Constant-Acceleration Residual Analysis. To directly assess whether generated motions obey Newtonian kinematics, we compute the mean discrete second-order residual $\phi_{t+1} - 2\,\phi_t + \phi_{t-1}$, averaged over all 32 frames of the sliding-down-ramp scenario for each method, as in Figure 5. This residual is zero for ideal constant-acceleration motion and therefore serves as a sensitive diagnostic of dynamical consistency. Figure 6 shows the horizontal (top) and vertical (bottom) residual fields. The SFT baseline and all PISA variants produce strong red/blue activations, indicating noticeable violations of the constant-acceleration constraint. Even methods that use ground-truth visual signals (PISA Depth, Segmentation, and Optical Flow) retain substantial structured residuals, revealing that pixel-level alignment does not translate into correct governing dynamics. In contrast, `NewtonRewards` produces markedly smoother residual maps with minimal magnitude, achieving the lowest absolute residuals across both axes. These reductions demonstrate that enforcing Newtonian kinematic structure yields trajectories that more closely adhere to true constant-acceleration behavior, beyond what can be captured through appearance- or flow-based supervision alone.

6.3. Ablation Study

Newtonian Kinematic Residual Constraint. Ablating the discrete residual term largely removes the temporal regularization effect, leading to only marginal overall improvement over SFT (+0.25%). Without enforcing constant acceleration, the model produces visually coherent but physically inconsistent motion, i.e., slightly better spatial fidelity (IoU +3.8%) yet degraded kinematic accuracy across L2, CD, velocity, and acceleration metrics. This highlights that the residual constraint is essential for stabilizing motion and aligning generated dynamics with Newtonian laws.

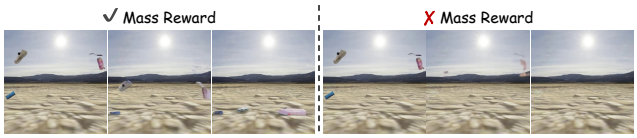

Figure 7. Videos generated with the mass reward (left) maintain consistent object persistence, while removing it (right) leads to degenerate behavior where objects vanish; an instance of reward hacking when optimizing only the kinematic residual.

Mass Conservation Reward. Although the overall gain slightly decreases (+8.1% vs. +9.75%) when removing mass conservation, as seen in Table 2, this configuration only employs the Newtonian kinematic residual without additional regularization. To understand its influence, we assess the role of mass conservation in stabilizing post-training.

> **Finding 2.** *The mass 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 mass consistency, the mass reward prevents this collapse and ensures stable, physically meaningful motion as in Figure 7.

7. 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 optical flow and visual features as surrogates for velocity and mass, `NewtonRewards` enforces Newtonian dynamics through kinematic and mass-conservation constraints. Our method enables video generators to obey constant-acceleration dynamics and maintain physical plausibility across diverse NMPs. Beyond Newtonian mechanics, our approach is inherently general. This 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.

Acknowledgements

This research was supported by NSF grant IIS-2212046, ANR-22-CE23-0007, ANR-22-CE39-0016, Hi!Paris grant and fellowship, DATAIA Convergence Institute as part of the "Programme d'Investissement d'Avenir" (ANR-17-CONV-0003) operated by Ecole Polytechnique, IP Paris, and was granted access to the IDRIS High-Performance Computing (HPC) resources under the allocation 2025-AD011015894 made by GENCI.

References

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

[2] 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. 6

[3] 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. 1

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025. 1

[5] 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*, pages 1–11, 2024. 2

[6] 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. 2

[7] 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*, 2023. 1, 2

[8] 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*, pages 22563–22575, 2023.

[9] 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, 2024. 2

[10] 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. 1

[11] 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. 2, 3

[12] Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Kalogeiton Vicky. Et the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *ECCV*, 2024. 1

[13] 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)*, pages 2553–2560. IEEE, 2022. 5

[14] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025. 1

[15] 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. 1

[16] 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. 1, 2

[17] 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. 2

[18] 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*, pages 3749–3761, 2022. 3, 5

[19] 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. 2

[20] 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*, pages 330–348. Springer, 2024. 2

[21] 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. 1

[22] 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. 1, 2

[23] 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*, 2022. 2

[24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646, 2022. 2

[25] 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. 2

[26] 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. 1, 2

[27] 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. 1

[28] 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. 2, 3

[29] 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. 1

[30] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning*, 2025. 1, 2

[31] 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. 1, 2

[32] 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. 2

[33] 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. 1

[34] 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*, pages 34–50. Springer, 2022. 2

[35] 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*, 2025. 1

[36] 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*, 2025. 1, 2, 3, 5, 6, 7

[37] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 2

[38] 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. 1

[39] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *ECCV*, 2024. 1, 2

[40] Jiaxi Lv, Yi Huang Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu 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. 1, 2

[41] 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. 2

[42] 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*, 2025. 1, 2

[43] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. 1, 2

[44] Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025. 2

[45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023. 2

[46] 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. 2

[47] 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. 5, 6

[48] 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. 1

[49] 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. 2

[50] 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. 2

[51] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 6

[52] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2

[53] 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*, 2025. 1, 2

[54] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025. 1, 2, 3

[55] Xi Wang, Robin Courant, Marc Christie, and Kalogeiton Vicky. Akira: Augmentation kit on rays for optical video generation. In *CVPR*, 2025. 1, 2

[56] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025. 2

[57] 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. 2

[58] 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. 2

[59] 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. 2

[60] Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. *arXiv preprint arXiv:2505.13934*, 2025. 1

[61] 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*, pages 10793–10804, 2025. 1, 2

[62] 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. 1, 2

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024. 6

[64] 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. 1

[65] 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. 1, 2

[66] 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*, 2024. 1, 2

[67] 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. 1

[68] 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. 2

[69] Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. *arXiv preprint arXiv:2510.11512*, 2025. 1

[70] Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *arXiv preprint arXiv:2509.21309*, 2025. 1, 2

[71] 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*, 2025. 1

[72] 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*, 2024. 2

[73] Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025. 1, 2

[74] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025. 1, 2

[75] 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*, 2024. 2

[76] 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. 1, 3

[77] 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. 1, 2, 6

What about gravity in video generation? Post-Training Newton's Laws with Verifiable Rewards

Supplementary Material

This supplementary document provides additional results and analyses supporting the main paper. Section A presents a *real-world evaluation* on 361 free-fall videos from the PISA [36] benchmark, showing that our physics-grounded post-training–developed entirely in simulation–transfers to natural scenes and real gravitational motion. Section B provides *additional quantitative evidence* for Finding 2, demonstrating mass conservation reward prevents reward hacking by avoiding degenerate zero-motion solutions. Section C offers *extended qualitative comparisons* across the remaining NMPs, with frame-level visualizations highlighting the improved consistency and stability of `NewtonRewards`. We also include *video-based qualitative results*: comparisons for all NMPs and real-world sequences at multiple frame rates (16 fps, 8 fps, 4 fps), and the teaser video from the main paper (see attached videos).

A. 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 [36]. 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 8). This setting allows us to assess whether a model trained purely on synthetic physics signals can generalize to real camera imagery and real-world dynamics.

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 ($\mathrm{RMSE}_{\mathbf{v}}$, $\mathrm{RMSE}_{\mathbf{a}}$). 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.

B. Reward Hacking Mitigation

In Section 6 of the main paper, we identified *reward hacking* as a failure mode that arises when optimizing only the kinematic residual. Without the mass-conservation reward, the generator can trivially reduce $\left\|\phi_{t+1} - 2\,\phi_t + \phi_{t-1}\right\|_2^2$ by driving all velocity fields ϕ_t toward zero–producing videos in which the object barely moves or even disappears. Here, we provide additional quantitative evidence of this finding.

Figure 8. **Real-world free-fall evaluation.** We test whether models post-trained purely in simulation can generalize to real camera imagery and real gravitational motion. Shown here is a representative video from the PISA real-world dataset (361 free-fall videos).

Table 4. **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 mass 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 mass	0.000986	**0.000331**	**0.033854**

As shown in Table 4, removing mass conservation (`NewtonRewards` w/o mass) 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 our claim that the mass-conservation reward is crucial for preventing reward hacking and stabilizing physics-grounded post-training.

C. Extended Qualitative Results

In addition to Fig. 5, we provide visual comparison in Figs. 9 to 12 between post-training strategies on our `NewtonBench-60K` for the remaining NMPs.

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 ($\mathrm{RMSE_v}$, $\mathrm{RMSE_a}$) 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 [36] 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 [36] ORO Depth Map	0.1704 (+0.70%)	0.4342 (+1.00%)	0.0218 (+10.10%)	2.4395 (-0.37%)	18.3474 (-0.38%)	+2.21%
PISA [36] ORO Segmentation	0.1712 (+0.23%)	0.4372 (+0.32%)	0.0210 (+6.06%)	2.4273 (+0.87%)	18.2344 (+0.99%)	+1.29%
`NewtonRewards`	**0.1698** (+1.05%)	**0.4333** (+1.21%)	**0.0235** (+18.69%)	**2.3889** (+2.43%)	**18.1670** (+1.36%)	**+4.15%**

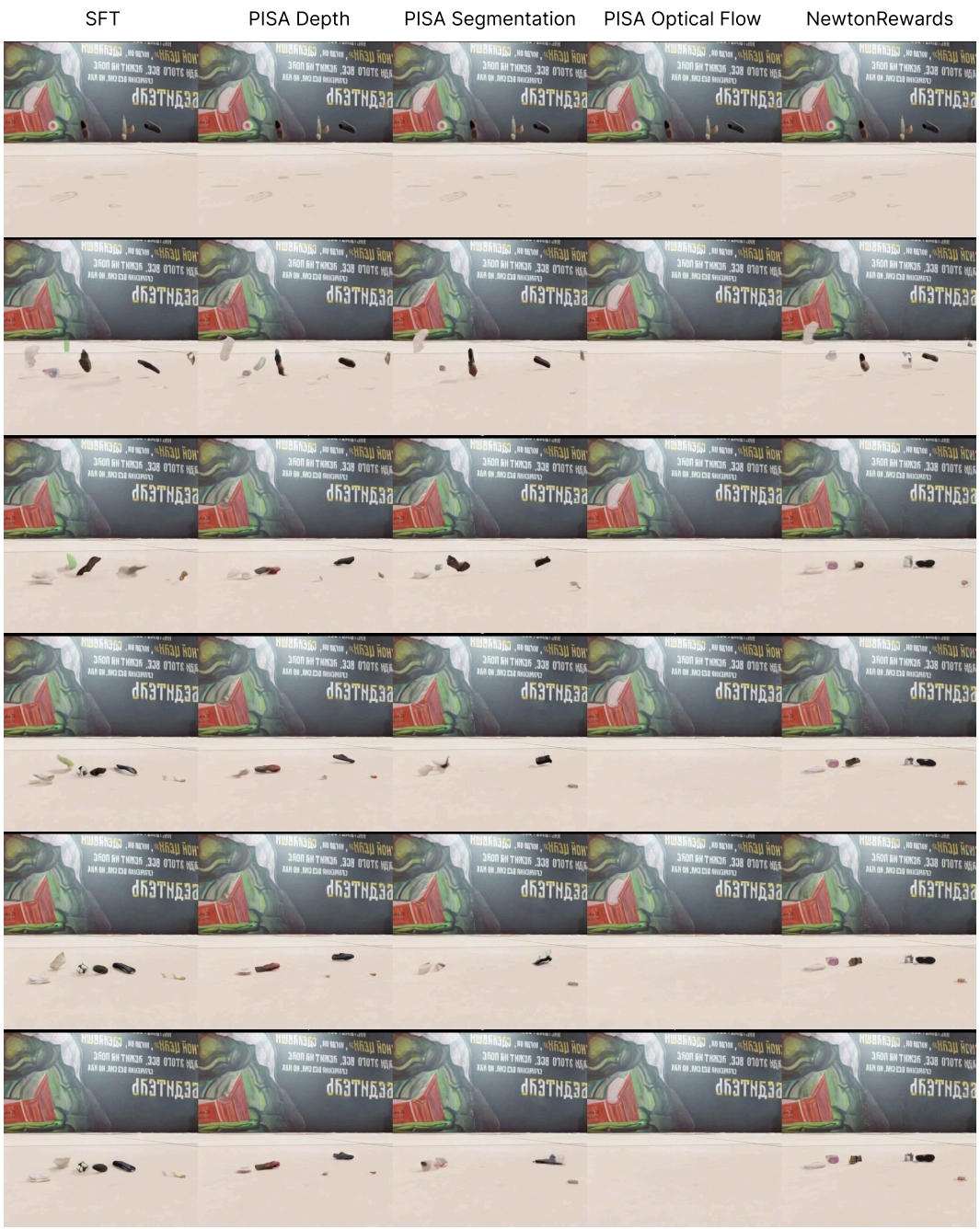

Figure 9. **Qualitative comparison of post-training strategies on the `NewtonBench-60K` free fall scenario.** Clear differences appear when examining the temporal evolution across frames (top→bottom). Under SFT and all PISA variants (Depth, Segmentation, Optical Flow), the objects frequently display inconsistent vertical acceleration and unstable trajectories sequence frames where items jitter, deviate laterally, or momentarily "hover". PISA Optical Flow particularly exhibits frame-to-frame jitter and irregular descent. In contrast, NewtonRewards produces smooth, stable free-fall motion: objects drop along physically plausible vertical paths with consistent acceleration and minimal horizontal drift, closely matching true gravitational dynamics.

Figure 10. **Qualitative comparison of post-training strategies on the `NewtonBench-60K` ramp slide up scenario.** Across temporal progression (top→bottom), SFT and all PISA variants (Depth, Segmentation, Optical Flow) exhibit inconsistent contact dynamics. In SFT, PISA Depth, and PISA Segmentation, the objects are sliding oppositionally down; and in PISA Optical Flow the object just disappears. In contrast, NewtonRewards produces coherent, physically grounded motion: the objects sliding up smoothly, with realistic frictional motion, and no frame-to-frame jitter. The resulting trajectory aligns closely with expected dynamics under gravity and surface friction.

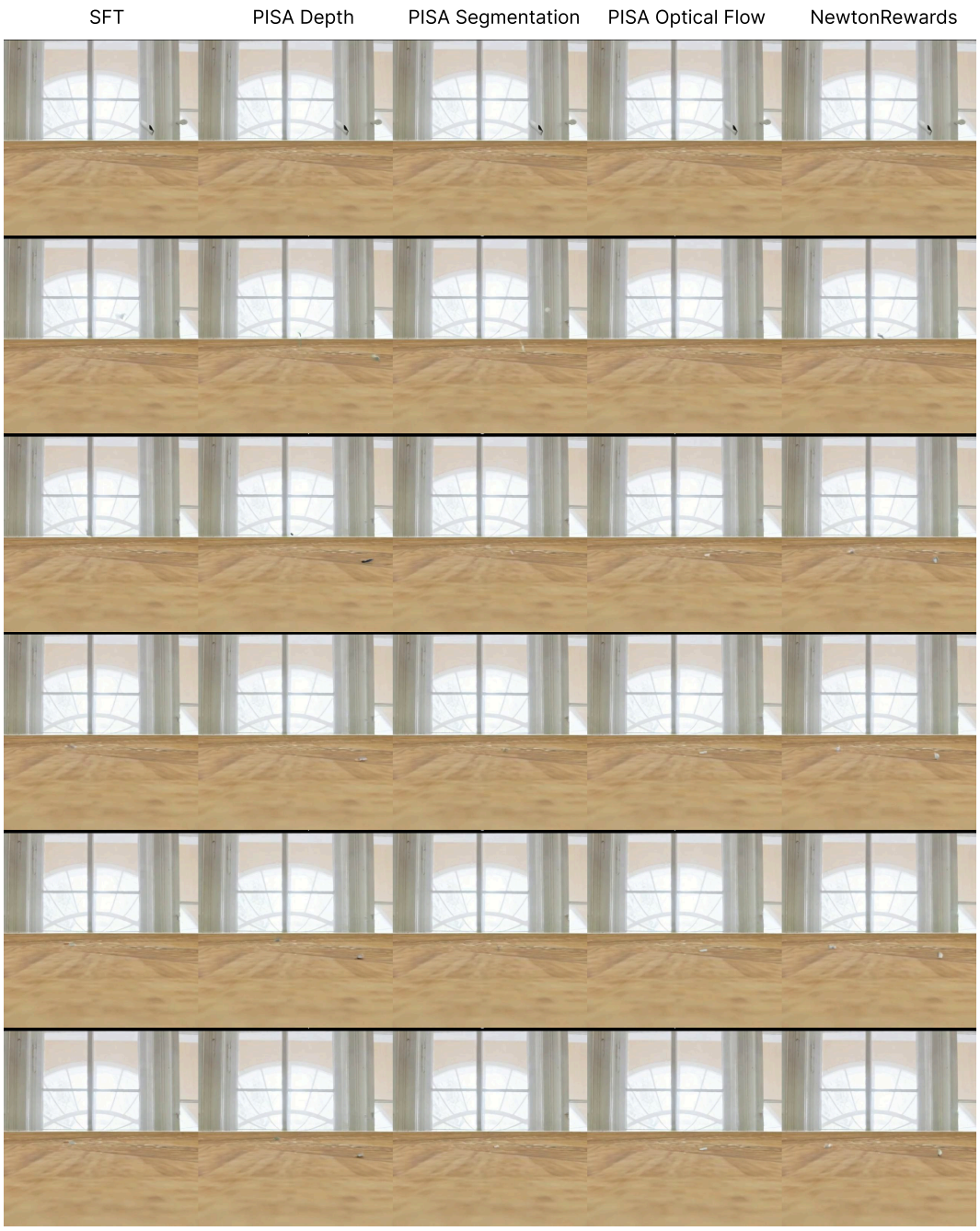

Figure 11. **Qualitative comparison of post-training strategies on the `NewtonBench-60K` horizontal throw scenario.** Examining the temporal rollout (top→bottom), SFT and all PISA variants (Depth, Segmentation, Optical Flow) exhibit inconsistent motion: objects either lose horizontal velocity too quickly, drift irregularly, or jitter frame-to-frame. Several PISA variants show abrupt slowdowns or curved, non-ballistic paths. In contrast, NewtonRewards produces smooth, coherent trajectories that follow a realistic horizontal-throw profile: constant horizontal velocity, stable parabolic descent, and no unnatural jitter. The motion aligns closely with classical projectile dynamics, demonstrating significantly improved physical fidelity.

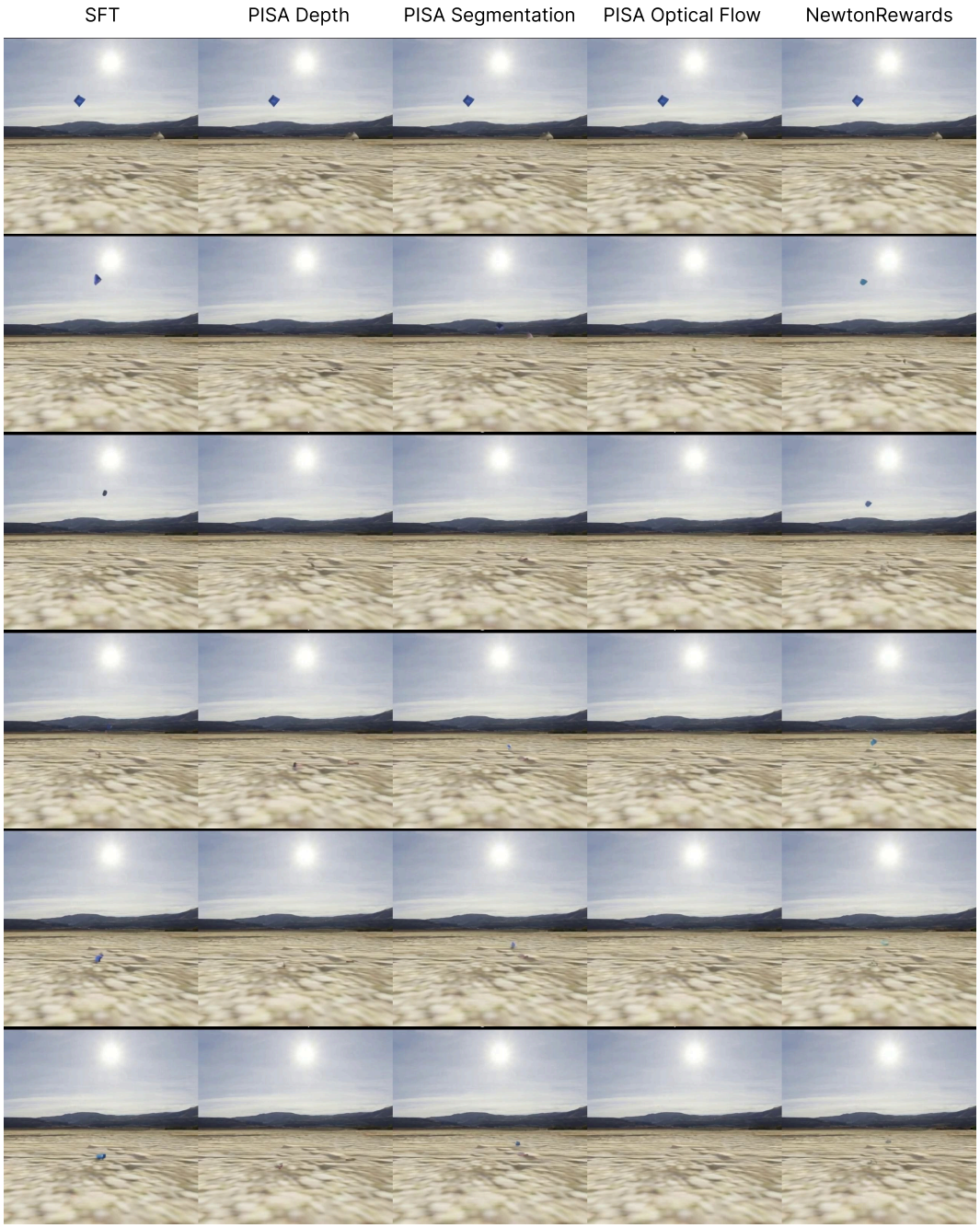

Figure 12. **Qualitative comparison of post-training strategies on the `NewtonBench-60K` parabolic throw scenario.** Across the temporal rollout (top→bottom), SFT and all PISA variants (Depth, Segmentation, Optical Flow) struggle to reproduce coherent parabolic motion. The thrown object exhibits noticeable inconsistencies–trajectory, abrupt velocity changes, or overly flattened arcs. In particular, objects disappear for the PISA Optical Flow case. In contrast, NewtonRewards generates smooth, physically realistic motion: the object follows a stable parabolic path with consistent horizontal velocity and gravitationally governed vertical acceleration.