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

Objects in Generated Videos Are Slower Than They Appear: Models Suffer Sub-Earth Gravity and Don't Know Galileo's Principle...for now

Varun Varma Thozhiyoor[1] Shivam Tripathi[1] Venkatesh Babu Radhakrishnan[1] Anand Bhattad[2]

[1]Indian Institute of Science [2]Johns Hopkins University

Project: https://gravity-eval.github.io

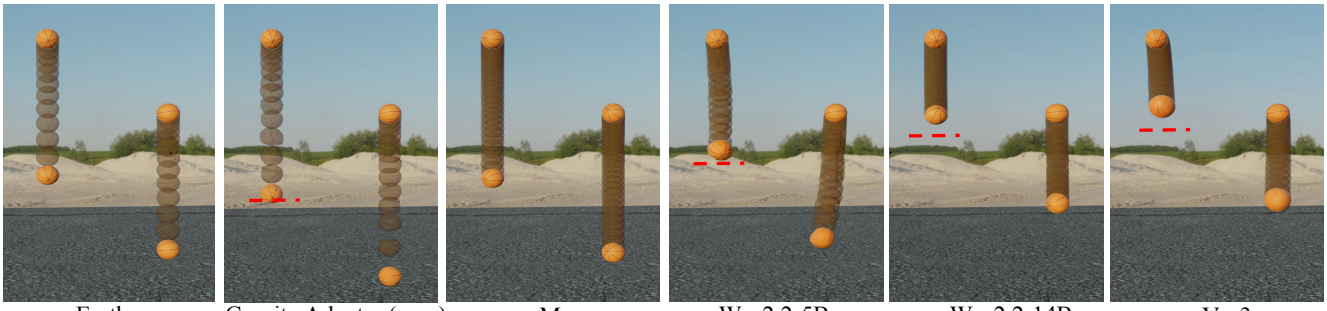


Figure 1. **Video generators produce physically implausible slow-motion falls and most fail to understand that objects fall at equal rates.** We visualize two identical balls dropped simultaneously from different heights using stroboscopic (time-lapse) composites, tracking motion until the lower ball impacts the ground. **Two failures emerge:** (1) **Galileo's principle violations:** Under Galileo's principle, both balls should fall equal distances in equal time regardless of starting height. Red dashed lines mark the expected position of the higher ball if both fell at equal rates. Earth (leftmost; simulation from Blender) shows the higher ball reaching this expected position, confirming correct physics. Moon reference ($g \approx 1.6\,\mathrm{m/s^2}$) shows both balls falling slower but preserving equal-rate progression. In contrast, Wan 14B and Veo3 show the higher ball severely lagging behind the red line. It traveled far less distance than the lower ball despite falling for the same time, violating the fundamental principle that gravitational acceleration is universal. Even Wan 5B shows noticeable lag. (2) **Severe under-acceleration:** The spacing between successive ball positions indicates effective acceleration; wider spacing means higher acceleration. Most models exhibit compressed spacing comparable to Moon ($1.6\,\mathrm{m/s^2}$) or Mars ($3.7\,\mathrm{m/s^2}$) rather than Earth's $9.81\,\mathrm{m/s^2}$, revealing motion dramatically slower than terrestrial physics. Our Gravity Adapter (second panel), fine-tuned on Wan 5B with just 100 examples, corrects both failures by bringing the higher ball to the expected position and improving Earth-like spacing.

Abstract

Video generators are increasingly evaluated as potential world models, which requires them to understand physical laws. We investigate their representation of a fundamental law: gravity. Out-of-the-box video generators consistently produce objects falling under an effectively lower acceleration. However, these physical tests are often confounded by an ambiguous metric scale. We first investigate whether the observed physical errors are artifacts of these ambiguities (e.g., incorrect assumptions about frame rate). We find that even temporal rescaling cannot correct the high-variance gravity artifacts. To rigorously isolate the underlying physical representation from these confounds, we introduce a unit-free, two-object protocol that tests the timing ratio $t_1^2/t_2^2 = h_1/h_2$, a relationship independent of g, focal length, and scale. This relative test reveals violations of Galileo's equivalence principle. We then demonstrate that this physical gap can be partially mitigated with targeted specialization. A lightweight low-rank adaptor fine-tuned on only 100 single-ball clips raises g_{eff} from $1.81\,\mathrm{m/s^2}$ to $6.43\,\mathrm{m/s^2}$ (reaching 65% of terrestrial gravity). This specialist adaptor also generalizes zero-shot to two-ball drops and inclined planes, offering initial evidence that specific physical laws can be corrected with minimal data.

1. Introduction

Two balls, a basketball and a tennis ball, are dropped from the same height, side by side. Assuming no air resistance, which hits the ground first? The answer is neither: they strike the ground simultaneously. Galileo demonstrated this principle over 400 years ago, and it remains one of the most fundamental predictions of Newtonian mechanics [9, 25]. In this paper, we investigate whether state-of-the-art video generators capture the fundamental physical law of gravity.

Galileo's principle is qualitative. A more rigorous, quantitative test is *how* objects fall: the time to impact is proportional to the square root of the height ($t \propto \sqrt{h}$). But testing this in a generative model is not straightforward. Recent work shows that while models learn rich 3D representations [4, 10, 12, 37], this internal geometry is inherently scale-ambiguous, and video frame rates provide no absolute time reference. Any rigorous test of physical law must therefore eliminate these confounds.

Prior benchmarks audit physics through trajectory fitting [20], conservation laws [38], or broad qualitative suites [2, 18, 24, 34]. However, these approaches either require camera calibration, rely on approximate heuristics, or optimize for absolute positional accuracy in a calibrated space – conflating the law of gravity with its parameterization. But none of these methods directly measures whether a model's representation of gravity quantitatively matches Earth's $g = 9.81\,\mathrm{m/s^2}$ in a way that is robust to the scale and time-base ambiguities inherent in generated video.

We introduce a unit-agnostic protocol based on relative timing between simultaneously falling objects. Because generated videos provide *no* guarantee that the model's internal 'meter' aligns with physical units or that frame intervals correspond to absolute time, we must eliminate reliance on explicit units. Under a static pinhole camera with zero initial velocity, the ratio of impact times satisfies

$$\frac{t_1^2}{t_2^2} = \frac{h_1}{h_2},$$

where metric scale, focal length, time base, and gravitational constant all cancel—yielding a calibration-free diagnostic of physical understanding.

We find that, out of the box, most models violate Galileo's principle that objects fall at the same rate; see Figure 1. In our single-ball drop experiments ($h \in [0.5, 4.0]\,\mathrm{m}$), impacts are consistently late across all tested video generators (Wan 5B, Wan 14B, Veo3, Cosmos 2B, Cosmos 14B). Interpreted in their native time gauge, the implied effective gravity g_{eff} ranges from 0.38 to $2.27\,\mathrm{m/s^2}$, far below Earth's $9.81\,\mathrm{m/s^2}$; balls therefore fall much more slowly than real-world gravity would predict, exhibiting an effective "sub-Earth" gravity. A straightforward temporal rescaling can move the *mean* g_{eff} closer to $9.81\,\mathrm{m/s^2}$, but the resulting distributions remain broad with large variance and heavy tails, indicating substantial inconsistency across scenes. Moreover, our two-object protocol reveals systematic, nonzero time differences Δt between the two balls when they traverse the same vertical distance. This relative-timing discrepancy is invariant to any global time rescaling and thus indicates a genuine violation of Galileo's principle. Even more surprisingly, larger models (Wan 14B, Cosmos 14B) exhibit *slower* motion than their smaller counterparts (Wan 5B, Cosmos 2B), contradicting the assumption that scale alone improves physical consistency.

To repair this error, we train a lightweight LoRA adapter on just 100 single-ball sequences. Applied to Wan 5B, it increases g_{eff} on average from $1.81\mathrm{m/s^2}$ to $6.43\mathrm{m/s^2}$—reaching 65% of terrestrial gravity on average. Interestingly, this specialist generalizes zero-shot to two-ball drops and inclined planes despite never seeing these scenarios during LoRA training. In summary, objects in generated videos are slower than they appear. Video generators excel as generalists but struggle as physics engines; our adapter supplies targeted specialization with minimal data. Just as Galileo showed that the falling rate is independent of mass, we show that physical accuracy is independent of model scale; it requires targeted correction, not raw capacity.

Contributions. **(1)** We introduce a unit-free, two-object measurement protocol that isolates gravitational acceleration while being invariant to camera scale and frame rate, providing a rigorous diagnostic for physical consistency in video generators. **(2)** We quantify systematic physics violations across state-of-the-art models, revealing effective gravity at 5–20% of Earth's value in their native time gauge and showing that larger models can be *less* physically accurate than smaller ones. **(3)** We show that a 100-example LoRA adapter on Wan 5B and Wan 14B partially improves these gravity metrics and shows generalization to unseen scenarios, demonstrating the feasibility of correcting specific physical laws with minimal data.

2. Related Works

Physics in Video Generators. Recent benchmarks evaluate physics understanding through pixel-to-ground-truth matching [20, 24], human/VLM-as-a-judge [2, 14, 15, 21], or trajectory modeling [18, 38]. However, these approaches face a fundamental challenge: scale and time ambiguity. Monocular vision [11, 40] and neural rendering [1] cannot recover absolute metric scale without calibration. Video generators inherit this limitation: their internal representations are unitless, and frame rates provide no absolute time base [16]. Yet existing physics benchmarks implicitly assume calibrated measurements, conflating the physical law itself with its parameterization (e.g., testing $g = 9.81\,\mathrm{m/s^2}$ specifically rather than the square-root scaling law). While VLMs enable broad qualitative evaluation [13, 23], they suffer from hallucinations and prompt shortcuts [23, 28], making them unreliable for quantitative physical laws. We address this by introducing unit-free relative measurements.

Falling Objects. For free-fall specifically, PISA [20] optimizes trajectory IoU using 5,000 simulated videos with reward-based fine-tuning, while Morpheus [38] compares extracted trajectories against physics simulations. However, neither quantifies whether models understand gravity's fundamental properties: the square-root time-height relationship ($t \propto \sqrt{h}$) or Galileo's equivalence principle (all objects fall at equal rates). They only measure deviation

from ground-truth videos with implicit, unspecified gravity values. Kang et al. [18] train 2D models from scratch to test generalizability but similarly do not isolate the physical law from scale ambiguity. Our two-ball protocol eliminates these confounds by testing timing ratios ($t_1^2/t_2^2 = h_1/h_2$), where scale, focal length, and gravity all cancel. This provides a calibration-free test of gravity.

Probing Knowledge in Generative Models. Recent work investigates what visual knowledge generative models encode, including depth, normals, semantics, and object relationships [6, 12, 37], as well as disentangled representations of lighting and geometry [4, 5, 10, 35]. Sarkar et al. [29] revealed that diffusion models systematically fail at projective geometry, mispredicting shadows and vanishing points despite encoding 3D cues. Our work extends this inquiry from static geometry to temporal dynamics, uncovering systematic failures in gravitational physics. Critically, we find that a 100-example LoRA corrects these failures and generalizes zero-shot to unseen scenarios, suggesting physical laws occupy sparse, learnable subspaces within foundation models. This connects interpretability research with physics-guided generation, showing that models encode latent physical knowledge that requires targeted activation rather than learning from scratch.

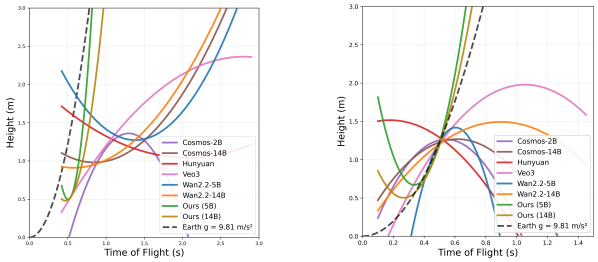
Physics-Guided Generation. Methods for improving physical correctness follow two paradigms: distilling physics through post-training alignment [20, 39] or explicitly simulating scenes with physics engines [7, 22, 33, 36]. Post-training approaches like PISA require massive compute, while simulation-based methods depend on accurate scene estimation and cannot handle complex real-world scenarios. Both align with the intuitive physics hypothesis [3] that physical understanding emerges from observing dynamics. However, our findings reveal that despite exposure to vast video data, models learn correlational patterns without internalizing fundamental rules. They systematically violate both acceleration laws and Galileo's equivalence principle. Unlike prior work, our lightweight LoRA adapter achieves substantial correction and generalizes zero-shot to two-ball drops and inclined planes, demonstrating that targeted specialization can efficiently bridge the gap between visual plausibility and physical correctness.

3. Experimental Setup

We investigate how well modern video generators capture gravitational physics in the canonical setting of falling objects. Our evaluation requires controlled generation and measurement; we describe our synthetic benchmark, the models tested, and our tracking-based evaluation protocol.

3.1. Synthetic Benchmark

Rendering. Each video in our benchmark is created synthetically using a procedural Blender pipeline[8]. We use various standard sports balls as falling objects, which should be well-represented in the models' training data. Balls are simulated to be dropped from random heights in the range $[0.5, 4.0]\,\mathrm{m}$, and the initial height is recorded. Videos are rendered at 1280×720 resolution for 2 seconds at 24 fps, yielding 48 frames per sequence.



(a) Original (b) Mean Scaling

Figure 2. **Effect of time-scaling on h–t relationships. (a)** We plot h versus t for all models. We repeat each test example with 4 seeds and fit polynomials through the means. The gray dashed line indicates terrestrial motion. All models systematically under-accelerate, and none obey the square root scaling law of time with height. The Gravity Adapters (green, gold) substantially improves Wan 5B and Wan 14B towards correct gravity. **(b) Mean time scaling.** We compute a Mean time scalar using a subset of random 30 samples from our dataset, which scales the effective time of the 30 samples to better match the ground truth time. The Mean time scalar, when applied to the second subset of 45 samples, brings the mean effective gravity closer to 9.81, m/s^2 for many models, but the variance remains high indicating that under-acceleration is not simply a frame rate artifact.

We generate N videos with heights uniformly sampled from $[0.5, 4.0]\,\mathrm{m}$. The dataset includes K different ball types (basketball, soccer ball, tennis ball, volleyball, baseball) across M unique HDRI environments. For the adapter training (Sec. 5), we use 100 single-ball sequences; all other experiments use held-out test data.

The camera is fixed at a sufficient distance and height to capture the entire trajectory of the ball and is positioned perpendicular to the falling motion. The scene is lit, and the background is occupied by HDRI maps. We use indoor and outdoor HDRI maps with a variety of different ground materials to match the setting. Backgrounds are chosen to match the scale of the ball—objects in the background do not appear too large or too small compared to the ball. We include samples with backgrounds having objects, to provide more context, as well as samples with more plain backgrounds that provide less context for more challenging generations(additional objects provide scale context).

The camera uses a focal length of $f = 50, \mathrm{mm}$ and is positioned between 1.0 and 8.0 meters from the drop zone. Its height is set to half the drop height, with an additional offset sampled from the range $[-0.5, , 0.5]$ meters. This configuration ensures that perspective distortion remains minimal across the falling trajectory.

Table 1. **Effect of time-scaling on g_{eff} estimation across models. (a) Original.** We report effective gravity values computed as $g_{\text{eff}} = 2h/t^2$ (m/s^2). The ground truth is 9.81 m/s^2. All models under-accelerate, and Gravity Adapters consistently reduce this deficit. Reported mean values are averaged over four random seeds and all test examples. Median and Range values are across all seeds and test samples. **(b) Mean time scaling.** A global scalar (MTS) is estimated from a 30-sample subset and applied to a disjoint 45-sample split. This shifts several models closer to 9.81 m/s^2, but variance remains high.

(a) Original

Model	Mean ($\mathrm{m/s^2}$)	Median ($\mathrm{m/s^2}$)	Range ($\mathrm{m/s^2}$)
Cosmos 2B [31]	1.85	1.30	[0.23, 14.18]
Cosmos 14B [31]	1.51	1.01	[0.24,10.31]
Hunyuan [19]	1.97	1.15	[0.23,15.84]
Veo3 [27]	2.27	2.08	[0.28, 6.66]
Wan 5B [32]	1.81	1.24	[0.26,8.26]
Wan 14B [32]	2.18	1.19	[0.27, 59.98]
Gravity Adapter 5B [17]	6.43	6.38	[1.24, 16.64]
Gravity Adapter 14B [17]	5.51	5.63	[1.52, 11.67]

(b) Mean-Time Scaled

Model	MTS	Mean($\mathrm{m/s^2}$)	Median ($\mathrm{m/s^2}$)	Range ($\mathrm{m/s^2}$)
Cosmos 2B [31]	2.43	10.34	7.24	[1.40, 69.96]
Cosmos 14B [31]	2.77	10.86	7.19	[1.81, 76.04]
Hunyuan [19]	2.63	13.85	7.06	[1.55, 104.64]
Veo3 [27]	2.12	9.39	8.5	[1.26, 29.11]
Wan 5B [32]	2.43	10.24	7.22	[1.76, 49.15]
Wan 14B [32]	2.56	13.78	7.78	[1.75, 321.30]
Gravity Adapter 5B [17]	1.28	10.27	9.74	[2.4, 26.67]
Gravity Adapter 14B [17]	1.36	10.00	10.10	[2.82, 19.08]

Single-Ball Protocol We first quantify the effective gravitational acceleration that models implicitly use by measuring fall time vs. height. This establishes a baseline before we eliminate scale ambiguity with our unit-free two-ball protocol. We generate 75 sequences with drop heights uniformly sampled from $[0.5, 4.0]\,\mathrm{m}$. This tests whether models produce trajectories consistent with $t = \sqrt{2h/g}$, allowing us to measure effective gravity $g_{\text{eff}} = 2h/t^2$.

Two-Ball Protocol. To eliminate metric scale and time-base ambiguity, we generate 50 sequences with two (identical) balls dropped simultaneously from different heights h_1, h_2 within the same frame. To ensure both balls experience identical perspective effects, they are positioned at the same distance from the camera and separated horizontally by ball-diameter with an offset of 0.5 meters to prevent occlusion. This guarantees that any scale ambiguity affects both balls equally, making the timing ratio truly unit-free. Height ratios h_1/h_2 range from 0.25 to 3.5. The relative timing should satisfy $t_1^2/t_2^2 = h_1/h_2$, which cancels gravity, scale, focal length, and frame rate—providing a unit-free test of physical understanding.

3.2. Models Evaluated

We evaluate models with Image-to-Video capability to constrain the generations to fall from a known height specified using a conditioning image. Experiments are conducted on Wan 5B and Wan 14B [32], Veo3 [27], Hunyuan [19], and Cosmos 2B and Cosmos 14B [31]. For each model, we supply the first frame of the video and a text prompt as initial conditions. We use consistent text prompts across models to isolate physics understanding from prompt sensitivity.

3.3. Measurement Protocol

Generated videos are resized and adjusted to 1280×720 and set to 24 fps, following [24]. We track the ball using SAM2 [26], initialized with Blender centroids. Impact time t is determined when: (1) the bottom of the ball drops below a point located one ball-radius (in pixels) above the ground threshold y_{ground}, and (2) vertical velocity drops below $\varepsilon = 1\,\text{pixels/frame}$. The ground threshold prevents false detections for hovering balls; the velocity threshold accounts for impact deformation. For single-ball drops, we compute $g_{\text{eff}} = 2h/t^2$ and report mean, median and range. For two-ball drops, we compute timing ratios t_1^2/t_2^2 and compare against the theoretical h_1/h_2. We manually verify extreme acceleration results above $50\mathrm{m/s^2}$ and exclude any samples that are found to be incorrectly evaluated because of SAM failures.

4. Results

4.1. Single-Ball Drops: Sub-Earth Gravity

Figure 2a plots h versus t for all the models. Under correct physics, points should lie on the curve $t^2 = 2h/g$. All the models deviate from this reference.

Table 1a summarizes effective gravity measurements. No model reproduces Earth-like acceleration. Smaller models show less extreme under-acceleration (closer to Earth gravity) than larger models, though all remain far below $g = 9.81\,\mathrm{m/s^2}$. Paradoxically, scaling increases capacity but does not improve physical accuracy. In the case of Wan 14B, even though it reports a higher mean value, the median and the corresponding distribution reveal that the majority of generations remain substantially under-accelerated. The increased mean is primarily driven by a small number of extreme samples. Our lightweight Gravity Adapter, fine-tuned on just 100 sequences (Sec. 5), improves Wan 5B from g_{eff}=1.81 to g_{eff}=6.43 (range: 1.24 –16.64$\mathrm{m/s^2}$), demonstrating that targeted specialization can partially correct gravity. Experiment with more detailed prompts is in the supplementary material.

Is under-acceleration a frame rate artifact? No. We first test whether temporal stretching explains the low gravity values. Using the single-ball dataset, we split 75 samples into two subsets (30 and 45 videos). From the first subset, we compute each model's average time-scaling factor (*MTS*) as mean(t_{eff}/t_{gt}). We then apply this global scaling factor to videos in the second subset ($1/MTS \times t_{eff}$), computing scaled gravity values g_{scaled}. Tab. 1b shows the results across samples and seeds. Since the absolute scale

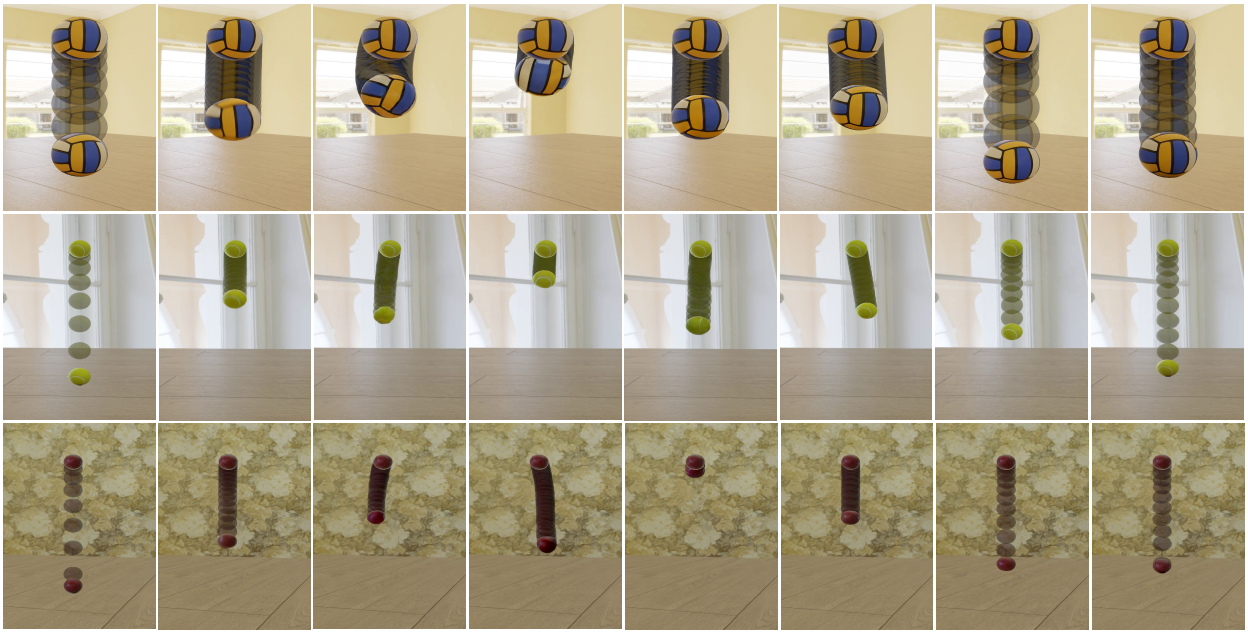


Figure 3. **Scaled Single-ball drops reveal systematic under-acceleration across all models.** Stroboscopic composites (left) visualize ball positions at equal time intervals from release. The panels show the trajectories performed by each model during the time it takes a ball falling under $9.8m/s^2$ to reach the ground, scaled by MTS (Tab. 1b). All models showcase severe under-acceleration (easily visible in the compressed spacing in the composites). The Gravity Adapters (seventh and eighth column) substantially improves Wan 5B and Wan14B toward terrestrial dynamics.

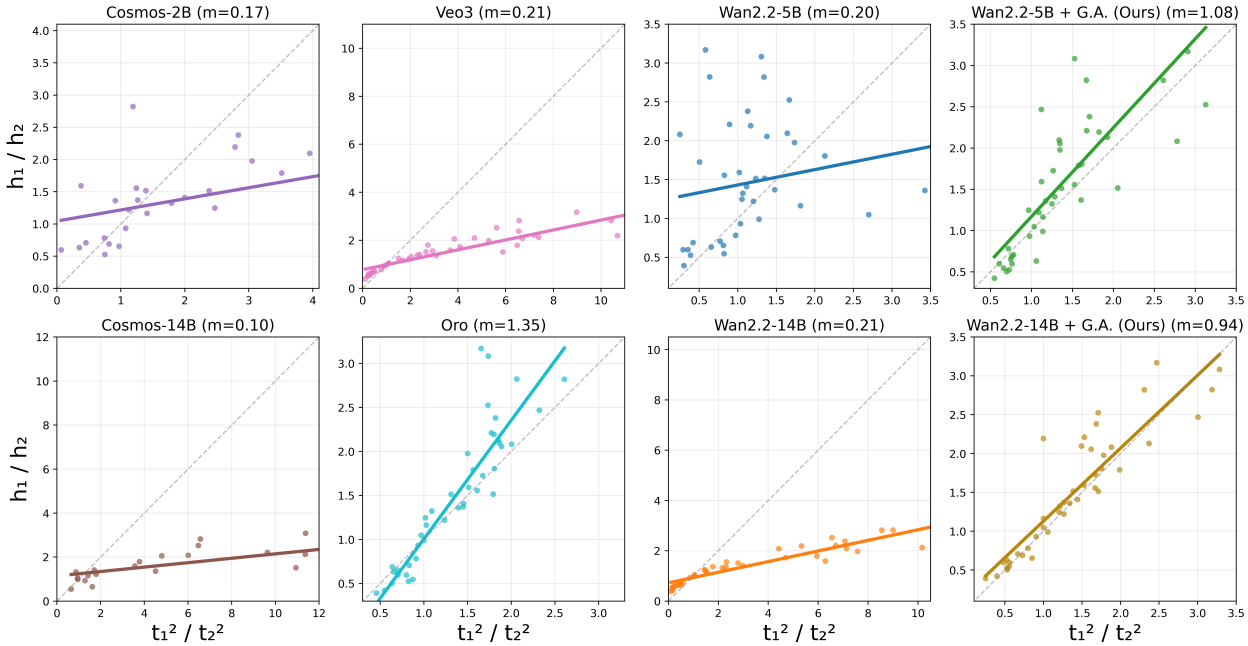


Figure 4. **Two-ball relative timing results.** We plot measured timing ratios t_1^2/t_2^2 against theoretical predictions h_1/h_2 across multiple height ratios. The gray dashed line indicates perfect agreement. All models deviate systematically, confirming that under-acceleration is not an artifact of scale estimation but reflects genuine physics and Galileo's principle violations. We also measure the slope (m) for each of them to understand the deviation.

Table 2. **Time deviation statistics (Δt in frames) for the two-ball falling experiment.** The lower ball hits the ground at time t_1 after traversing a vertical distance y_1. We then measure the time t_2 for the upper ball to traverse the *same* distance and report $\Delta t = t_2 - t_1$ (ideal constant-g motion gives $\Delta t = 0$). Positive values (e.g., Wan 14B, Veo3, Cosmos 14B) mean the upper ball is *slower*, indicating under-acceleration relative to the lower ball; negative values (Wan 5B, Cosmos 2B) mean it is *faster*, indicating over-acceleration. Interestingly, the *direction* of the violation flips with scale: smaller models tend to have negative mean Δt, while larger models have positive mean Δt. Gravity Adapters move the means close to zero (e.g., -0.95 for 5B and 0.14 for 14B), thereby shrinking the ranges, which largely cancels this scale-dependent Galileo violation.

Model	Mean (Δt)	Median (Δt)	Range (Δt)
Cosmos 2B [31]	-2.77	-2.0	[-35.0, 23.0]
Cosmos 14B [31]	4.03	5.0	[-36.0, 30.0]
Veo3 [27]	6.71	7.00	[-42.0, 28.0]
Wan 5B [32]	-4.22	-4.0	[-35.0, 30.0]
Wan 14B [32]	2.22	1.0	[-14.0, 16.0]
Gravity Adapter 5B [17]	-0.95	-1.00	[-8.0, 4.0]
Gravity Adapter 14B [17]	**0.14**	**0.0**	[-5.0, 3.0]

is unknown, we treat the single-ball experiment as a test of temporal consistency. If the model simply had a slow clock, a linear time scalar should fix the error. As Tab. 1b shows, it does not. This failure necessitates a unit-free metric to understand the root cause. See Fig. 3 for qualitative examples demonstrating under-acceleration. We also tested other scaling baselines (see supplementary material): (1) per-scene mean-time scaling computed across multiple seeds, and (2) height-adjusted scaling for non-vertical trajectories. Neither approach yields substantial improvement, confirming that physics errors persist even with privileged per-sample correction.

4.2. Two-Ball Drops: Failure of Galileo's Principle

Prior work evaluating physics in video generators[20, 38] predominantly tests single-object scenarios, comparing generated trajectories against ground truth. While such approaches can measure whether models follow text prompts accurately, they cannot determine whether models understand the underlying physical principles. A model might correctly match a single falling trajectory through a pattern recognition without understanding gravity.

We pose a more fundamental question: Do models understand Galileo's principle of gravitational equivalence – that all objects fall at the same rate regardless of mass or starting height? To test this, we generate scenes with two identical balls dropped simultaneously from different heights $h_1 < h_2$. We use identical balls to eliminate depth-perception confounds that could arise from size differences. Under correct physics, both balls experience identical acceleration and thus fall equal distances in equal time. At the moment when the lower ball impacts at time $t_1 = \sqrt{\frac{2h_1}{g}}$, the higher ball should have fallen the same distance $d = h_1$, placing it at height $h_2 - h_1$. **This is a zero-ambiguity test**: either both balls fall together (pass) or they don't (fail).

Figure 5 reveals that **most models catastrophically fail**. Veo3, Wan 14B, and Cosmos 14B show the higher ball traveling an unequal distance compared to the lower ball, when the lower ball lands. This observation is confirmed by plotting t_1^2/t_2^2 versus h_1/h_2 across multiple height ratios as shown in Figure 4. This is not just a uniform timing error; it reflects a failure to respect that gravitational acceleration is universal. Our two-ball experiments show that the upper and lower balls often take different times to traverse the same distance: in some models, the higher ball is effectively slower, in others faster, but in all cases the two objects experience different accelerations within the same scene, contradicting four centuries of physics.

To quantify deviations from Galileo's principle we measure the temporal lag between the two falling balls in units of video frames. First, we record the time at which the lower ball reaches the ground, denoted as t_1, having covered a vertical pixel distance y_1. We then measure the time t_2 taken by the upper ball to traverse the same distance, and compute the deviation $\Delta t = t_2 - t_1$. A positive deviation indicates that the upper ball requires more time to cover the same distance, reflecting under-acceleration relative to the lower ball; and a negative deviation implies the opposite, an over-acceleration. The measured deviations for each model are reported in Tab 2. We find an interesting pattern: the smaller models - Wan 5B and Cosmos 2B show an opposite trend compared to their larger counterparts, with the higher ball accelerating faster than the slower ball. As shown in Tab. 2. Our gravity adapters (Sec. 5) noticeably improve both Wan-5B and Wan-14B, bringing their behavior closer to the physically correct regime in which gravity acts universally, not separately on each object.

5. Gravity Adapter: Targeted Specialization

Video generators excel as generalists but struggle as physics engines. We investigate whether these physical deficits are due to a lack of model capacity or simply a lack of alignment. By training a lightweight adapter, we probe the learnability of gravitational physics.

5.1. Training Protocol

We train a LoRA [17, 30] on Wan 5B to correct its systematic under-acceleration. The training set consists of 100 single-ball drop sequences rendered in diverse HDRI environments distinct from our test benchmark. We train for 5,000 iterations with rank $r = 32$, learning rate 10^{-4}. Total training time is approximately 6 hours on two A100 GPUs.

Figure 5. **Most models fail Galileo's principle of gravitational equivalence.** We freeze at the moment the lower ball (from height h_1) impacts the ground. Under correct physics, both balls should have fallen equal distances in equal time, demonstrating that gravitational acceleration is universal (leftmost column: Ground Truth Earth). **Catastrophic failures:** Veo3 (sixth column) shows the higher ball barely moving while the lower ball lands—a complete violation of 400-year-old physics. Wan 14B, Veo3, and Cosmos 14B show similar failures with the higher ball remaining significantly elevated, suggesting these models believe in gravity depends on the starting height or object ordering. **Correction:** The Gravity Adapter (second column) finetuned on Wan 5B restores near-terrestrial acceleration *and* perfects equal-rate falling, demonstrating that this fundamental physics deficit can be corrected with only 100 training examples.

5.2. Results on Benchmark Tasks

Single-Ball Drops. Figure 2a shows that the adapter shifts the mean effective gravity significantly closer to Earth's standard ($1.81 \rightarrow 6.43 m/s^2$). While the distribution remains broad (indicating that stochastic generation artifacts persist), it shows that the model can be realigned to physical laws with negligible data (100 examples).

Two-Ball Drops (Zero-Shot). We also find that the adapter shows zero-shot transfer to two-ball relative timing tasks without ever seeing two-object scenarios during training (Figure 4). Timing ratios move substantially closer to the theoretical $t_1^2/t_2^2 = h_1/h_2$ prediction, indicating the adapter learned general gravitational dynamics rather than memorizing single-ball trajectories.

Real-World Transfer: PISA Benchmark. To test whether the adapter overfits to synthetic training data, we evaluate zero-shot on the PISA benchmark [20]—361 real-world videos of objects falling onto cluttered surfaces. This is challenging: our adapter was trained only on synthetic scenes with sports balls and clear ground planes. Table 3 shows the adapter outperforms the baseline on all metrics.

5.3. Comparison with Alternative Approaches

Table 4 compares our adaptor with alternative approaches.

Table 3. **Zero-shot generalization to real-world PISA benchmark.** The Gravity Adapter (5B), trained only on 100 synthetic sequences, improves trajectory accuracy on real-world videos [20]. The marginal improvement when full finetuning (FT) and reward optimization (ORO)[20] does not justify its huge training cost.

Model	L2 ↓	Chamfer ↓	IoU ↑	Params
Wan 5B (baseline) [32]	0.148	0.413	0.073	—
Wan 5B + Gravity Adapter [17]	0.127	0.346	0.086	**80M**
Wan 5B + FT [20] + ORO [20]	**0.123**	**0.332**	**0.089**	5B

LoRA Rank Ablation We increase the LoRA rank from 32 to 64 to examine whether additional capacity improves g_{eff}. Despite doubling the parameter count (80M $\rightarrow$ 160M), we observe no meaningful improvement, indicating that effective gravity cannot be enhanced through scaling alone. Lower-rank variants further degrade performance.

Full Model Fine-Tuning Replicating PISA's protocol [20], we fine-tune all Wan 5B parameters on our 100-example set using their two-stage approach: 5,000 supervised iterations followed by 1,000 reward-model iterations [26]. Due to computational constraints, we reduce resolution to 480p and 10 denoising steps. Our LoRA matches or exceeds full fine-tuning performance (Fig. 6) while training only 1% of parameters and using an order of magnitude less compute, suggesting physics correction may be

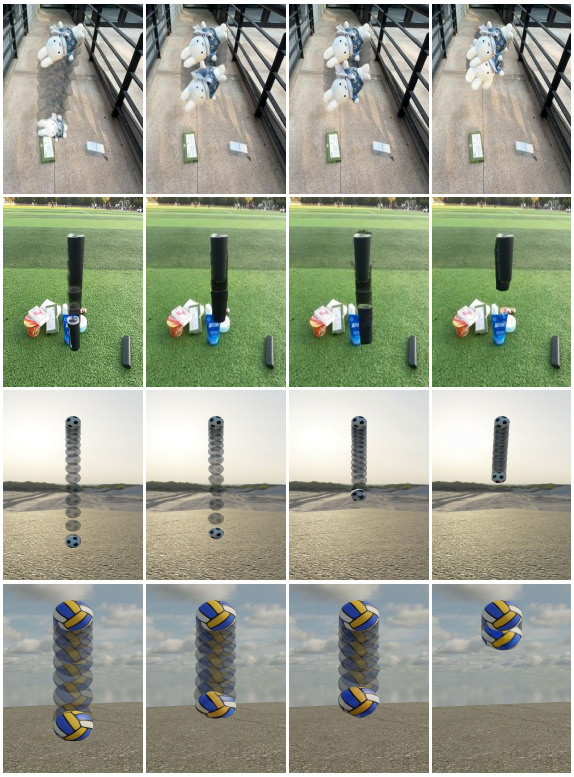


Figure 6. **Comparison with Full FT + ORO.** The Gravity Adapter (5B) shows impressive performance on both the real world dataset [20] and our benchmark with comparable and sometimes better performance at reduced computational cost.

Table 4. The Gravity Adapter (5B) exceeds full fine-tuning performance on our benchmark with less training computation

Method	**Mean** ($\mathrm{m/s^2}$)	**Median** ($\mathrm{m/s^2}$)	**Range** ($\mathrm{m/s^2}$)
Wan 5B (baseline) [32]	1.81	1.24	[0.26, 8.26]
Baseline + FT [20] + ORO [20]	4.07	3.63	[0.89, 10.72]
Gravity Adapter (ours, rank 8) [17]	3.07	2.83	[1.24, 16.64]
Gravity Adapter (ours, rank 16) [17]	5.64	5.72	[1.51, 12.05]
Gravity Adapter (ours, rank 32) [17]	**6.43**	**6.38**	[1.24, 16.64]
Gravity Adapter (ours, rank 64) [17]	6.11	5.72	[1.91, 14.17]

amenable to low-rank adaptation.

5.4. Generalization to Canonical Scenarios

To test whether the adapter learned general gravitational principles or merely ball-dropping heuristics, we evaluate on canonical physics regimes never seen during training:

Inclined Planes. We generate 12 sequences of smooth cubes sliding down frictionless inclines at angles from 30° to 75°. Under correct physics, acceleration should scale as $g\sin(\theta)$. Figure 7 shows the Gravity Adapter substantially improves motion realism compared to baseline Wan 5B, with objects accelerating appropriately for their incline angle. This demonstrates that the adapter internalized gravitational principles applicable beyond vertical free fall.

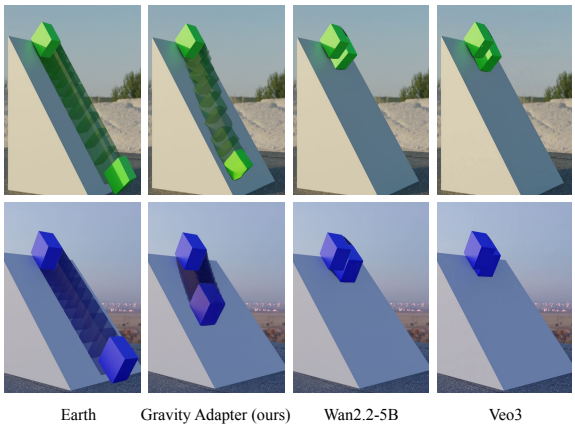


Figure 7. **Zero-shot generalization to inclined planes.** The Gravity Adapter (5B) improves motion realism for cubes sliding down inclines at various angles, despite never seeing this scenario during training. This demonstrates learning of general gravitational principles rather than ball-specific heuristics.

6. Discussion

We evaluate video generators on gravitational physics, revealing systematic failures in their native time gauge. Single-ball experiments show all models dramatically under-accelerate ($g_{\text{eff}} \sim 1$–$2\,\mathrm{m/s^2}$ vs. Earth's $9.81\,\mathrm{m/s^2}$), so balls fall much more slowly than real-world gravity would predict, exhibiting an effective "sub-Earth" gravity. A simple global temporal rescaling can move the *mean* g_{eff} closer to $9.81\,\mathrm{m/s^2}$, but the resulting distributions remain broad with large variance and heavy tails. Our unit-free two-ball protocol eliminates scale ambiguity by testing timing ratios ($t_1^2/t_2^2 = h_1/h_2$) and time deviations Δt for traversing the same distance. Most models catastrophically fail Galileo's principle, with different objects effectively experiencing different gravity within the same scene; these relative-timing violations are invariant to any global time rescaling. Surprisingly, larger models perform worse on these metrics, contradicting naive scaling-law expectations. A lightweight LoRA adapter trained on 100 synthetic examples substantially corrects these failures (e.g., $1.81 \rightarrow 6.43\,\mathrm{m/s^2}$) and moves Δt close to zero with tighter ranges, and it generalizes zero-shot to real-world videos, two-ball drops, and inclined planes.

Limitations and Future Work. Our adapter achieves only 65% of terrestrial gravity on average with high variance (1.24–$16.64\,\mathrm{m/s^2}$). We evaluate only vertical free-fall and simple inclined motion, not projectile trajectories, rotation, collisions, or friction. Future directions include: (1) Broader evaluation: Apply unit-free protocols to momentum, energy conservation, and fluid dynamics. (2) Physics-informed losses: Incorporate $d = \frac{1}{2}gt^2$ and relative-timing penalties directly into training objectives. (3) Multi-law adapters: Train single adapters for multiple physical phenomena.

Acknowledgements

This work is supported by a grant from Google and Kotak-IISc AI-ML Centre (KIAC).

References

[1] Joshua Ahn, Haochen Wang, Raymond A Yeh, and Greg Shakhnarovich. Alpha invariance: On inverse scaling between distance and volume density in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20396–20405, 2024. 2

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

[3] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110 (45):18327–18332, 2013. 3

[4] Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. Stylegan knows normal, depth, albedo, and more. *Advances in Neural Information Processing Systems*, 36: 73082–73103, 2023. 2, 3

[5] Anand Bhattad, James Soole, and David A Forsyth. Stylitgan: Image-based relighting via latent control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4231–4240, 2024. 3

[6] Anand Bhattad, Konpat Preechakul, and Alexei A Efros. Visual jenga: Discovering object dependencies via counterfactual inpainting. *arXiv preprint arXiv:2503.21770*, 2025. 3

[7] Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6178–6189, 2025. 3

[8] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 3

[9] Stillman Drake. *Galileo at work: His scientific biography*. Courier Corporation, 2003. 1

[10] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023. 2, 3

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2

[12] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D Awareness of Visual Foundation Models. In *CVPR*, 2024. 2, 3

[13] 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, 2025. 2

[14] Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. "phyworldbench": A comprehensive evaluation of physical realism in text-to-video models, 2025. 2

[15] Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation, 2025. 2

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4, 6, 7, 8

[18] 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 *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. 2, 3

[19] 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. 4

[20] 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. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. 2, 3, 6, 7, 8

[21] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models, 2025. 2

[22] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[23] Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility, 2025. 2

[24] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. 2, 4

[25] Isaac Newton, I Bernard Cohen, and Anne Whitman. *The Principia: mathematical principles of natural philosophy*. Univ of California Press, 1999. 1

[26] 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. *arXiv preprint arXiv:2408.00714*, 2024. 4, 7

[27] Google Research and DeepMind. Veo 3: Text-to-video model. Online model release, 2025. 4, 6

[28] Enes Sanli, Baris Sarper Tezcan, Aykut Erdem, and Erkut Erdem. Can your model separate yolks with a water bottle? benchmarking physical commonsense understanding in video generation models, 2025. 2

[29] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28140–28149, 2024. 3

[30] ModelScope Team. Diffsynth-studio: A diffusion model engine. https://github.com/modelscope/DiffSynth-Studio, 2025. GitHub repository, Apache-2.0 license. 6

[31] NVIDIA Cosmos Team. Cosmos-predict2: World foundation models for physical ai. https://github.com/nvidia-cosmos/cosmos-predict2, 2025. GitHub repository, Apache-2.0 license. 4, 6

[32] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4, 6, 7, 8

[33] 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. 3

[34] 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

[35] Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 442–452, 2025. 3

[36] 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. *arXiv preprint arXiv: 2509.21309*, 2025. 3

[37] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. *Advances in Neural Information Processing Systems*, 37:43468–43498, 2024. 2, 3

[38] Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments, 2025. 2, 6

[39] 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. 3

[40] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2