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

Supplementary Material

Table 1. **Use of expanded prompts.** Detailed prompts describing the scene with explicit parameters do not significantly affect the model's understanding of gravity. Wan 5B increase marginally in g_{eff}. Veo 3 and Cosmos 2B show a decline.

Model	Base Prompt			Expanded Prompt		
	Mean (m/s^2)	**Median** (m/s^2)	**Range** (m/s^2)	**Mean** (m/s^2)	**Median** (m/s^2)	**Range** (m/s^2)
Wan 5B	1.81	1.24	[0.26, 8.26]	2.06	1.37	[0.29, 6.80]
Veo3	2.27	2.08	[0.28, 6.66]	1.63	1.24	[0.34, 4.94]
Cosmos 2B	1.85	1.30	[0.23, 14.18]	1.25	0.83	[0.27, 3.84]

Table 2. **Explicit Guidance Models.** The Gravity Adapter (5B) outperforms methods with additional guidance.

Method	**Mean** (m/s^2)	**Median** (m/s^2)	**Range** (m/s^2)
Wan 5B (baseline) [6]	1.81	1.24	[0.26, 8.26]
Wan 5B + Gravity Adaptor [1]	**6.43**	**6.38**	[1.24, 16.64]
FLF (first + last frame) [6]	0.58	0.22	[0.03, 23.70]
Trajectory guided [7]	0.38	0.16	[0.04, 3.22]

1. Effect of Expanded text prompts

To reduce any scale ambiguity for the models, we experiment with more detailed text prompts containing explicit height, diameter of the ball, distance of the camera from the ball, and height of the camera above the ground. However, we see no significant improvement in any of the evaluated models (Tab 1).

2. Explicit Guidance Models

We test whether providing more information improves physical accuracy. The First-Last Frame model [6] receives initial and final frames to reduce depth ambiguity, but performs poorly—likely constrained to 5-second generations that default to slow motion. We also evaluate trajectory matching [7] using ground-truth 2D centroid trajectories from drop height to impact, after which we let the model freely generate. This model similarly underperforms ($g_{\text{eff}} = 0.38\,\text{m/s}^2$, range $0.04{-}3.22$), suggesting difficulty handling rapid motion. Our adapter exceeds both methods without any explicit guidance.

3. Additional Time Scaling Heuristics

Per Sample Time Scaling We also evaluate per-sample time scaling to account for seed-to-seed variance. For each sample, we use seeds 999 and 777 to compute an individual scaling factor, then apply it to seeds 42 and 123. Per-sample scaling yields higher g_{scaled} values (Tab. 3c) but the variance remains substantial, possibly due to high variance between seeds.

Adjusting height for angled trajectories. Many generated videos show non-vertical trajectories, which could result from the model interpreting the ground plane or camera as tilted. To provide the benefit of the doubt again, we adjust the effective drop height based on the observed angle of deviation, computing h_{adj} and the corresponding scaled gravity $g_{adj+scaled}$ using globally scaled times. Tab. 3d shows that this height adjustment provides no meaningful improvement. Even after applying perspective correction, the models still show substantial under-acceleration. The gravity distributions (Fig. 2b–2d) reveal that most samples remain below $9.81, \text{m/s}^2$, indicating that time scaling alone does not resolve the core physical error. Fig. 1b–1d shows the change in the h vs t plots when different time scaling techniques are applied.

4. Prompt Structure

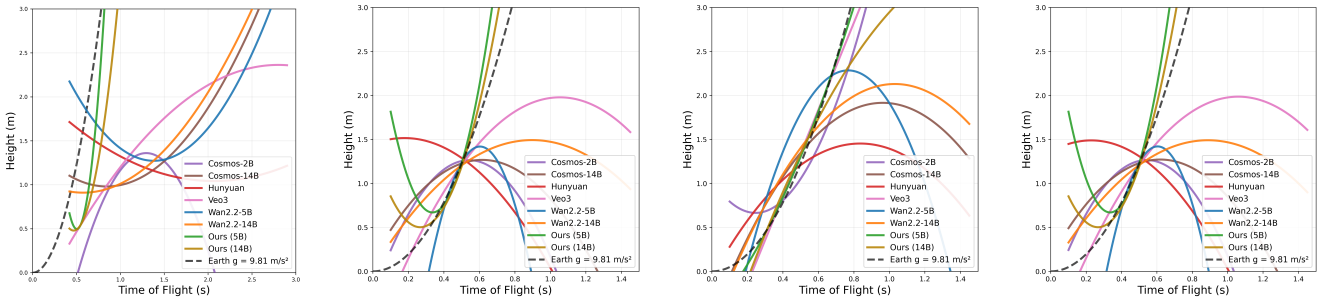
We use a consistent prompt structure across all experiments to isolate physics understanding from prompt sensitivity.

Single Ball Drops. A video showing a ball being dropped from a height onto the ground. The camera is static and positioned to clearly capture the vertical motion of the ball. The ball falls naturally under gravity, accelerating freely with no air resistance and hits the ground.

Two Ball Drops. A video showing two identical balls being dropped from two different heights onto the ground. The camera is static and positioned to clearly capture the vertical motion of both balls. Both balls fall naturally under gravity, accelerating freely with no air resistance, and hit the ground.

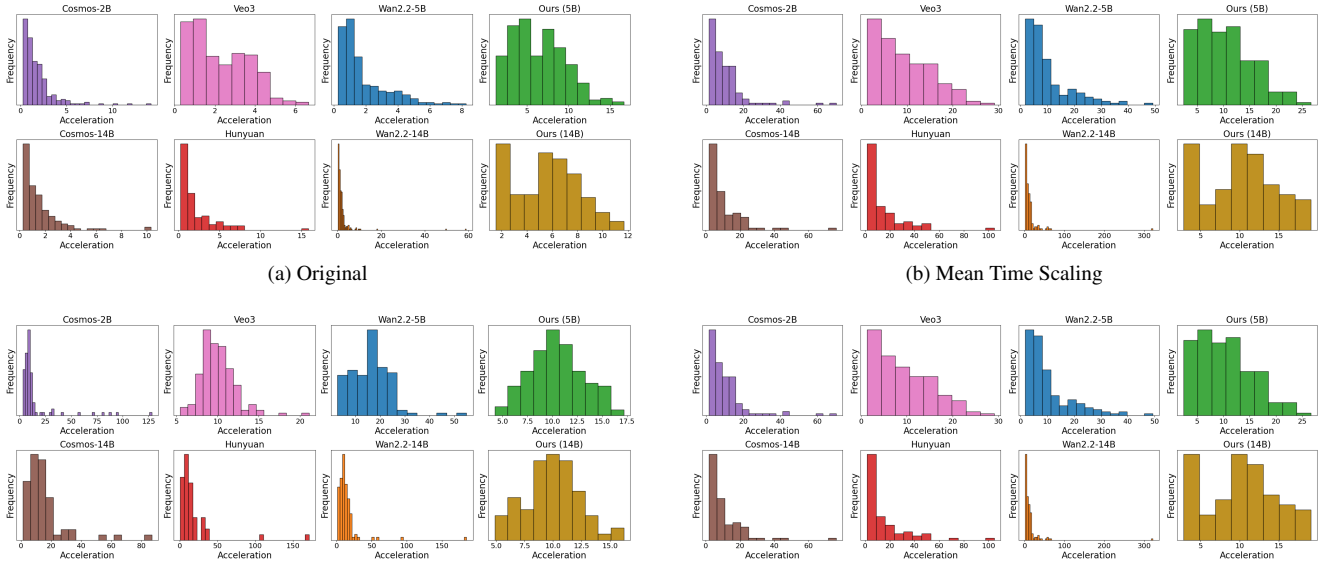
Expanded Prompt. A video showing a ball of diameter $diameter$ meters being dropped from a height of $initial\ height$ meters onto the ground. The camera is static and positioned at a height of $camera\ height$ meters, at a distance of $camera\ depth$ meters to clearly capture the vertical motion of the ball. The ball falls naturally under gravity at $9.8\ meters\ per\ second\ square$, accelerating freely with no air resistance, and hits the ground.

Inclined Plane. A video showing a smooth square block sliding down a frictionless inclined plane under gravity. The inclined plane is fixed and set at an angle, with no friction between the block and the surface. The block starts from rest near the top and accelerates uniformly as it slides



(a) Original (b) Mean Scaling (c) Per-Sample Scaling (d) Height-Adjusted Time Scaling

Figure 1. **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 close to 9.81, m/s2 for many models. **(c) Per-sample scaling.** Instead of a mean time scalar, we experiment with a per sample time scalar, computed using half of the seeds. The scalar is then applied to the other half of the seeds. **(d) Height-adjusted Mean Time scaling.** Accounting for deviation from the straight line vertical trajectory does not significantly improve the effective gravity.



(a) Original (b) Mean Time Scaling

(c) Per sample Time Scaling (d) Mean Time Scaling with Height adjustments

Figure 2. **Effect of Time scaling on distribution of gravities**

downward. The camera is static and positioned at the side to clearly capture the motion along the incline.

Table 3. **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. **(c) Per-sample time scaling.** Instead of a global correction, we compute a per-sample scalar using half the seed runs and apply it to the others. **(d) Height-adjusted scaling.** Accounting for deviations from ideal vertical motion does not meaningfully improve g_{eff} estimation, suggesting that model errors stem from deeper physics issues rather than trajectory shape.

(a) Original

Model	Mean (m/s^2)	Median (m/s^2)	Range (m/s^2)	Q1-Q3 (m/s^2)
Cosmos 2B [5]	**1.85**	1.30	[0.23, 14.18]	[0.68, 2.24]
Cosmos 14B [5]	1.51	1.01	[0.24,10.31]	[0.55, 1.92]
Hunyuan [2]	1.97	1.15	[0.23,15.84]	[0.48, 2.72]
Veo3 [4]	2.27	2.08	[0.28, 6.66]	[1.00, 3.38]
Wan 5B [6]	1.81	**1.24**	[0.26,8.26]	[0.77, 2.41]
Wan 14B [6]	2.18	1.19	[0.27, 59.98]	[0.63, 2.04]
Gravity Adapter 5B [1]	**6.43**	6.38	[1.24, 16.64]	[3.95, 8.65]
Gravity Adapter 14B [1]	**5.51**	5.63	[1.52, 11.67]	[3.02, 7.45]

(b) Mean-Time Scaled

Model	MTS	Mean(m/s^2)	Median (m/s^2)	Range (m/s^2)	Q1-Q3 (m/s^2)
Cosmos 2B [5]	2.43	10.34	7.24	[1.40, 69.96]	[3.74, 12.75]
Cosmos 14B [5]	2.77	10.86	7.19	[1.81, 76.04]	[4.02, 14.20]
Hunyuan [2]	2.63	13.85	7.06	[1.55, 104.64]	[3.23, 18.93]
Veo3 [4]	2.12	9.39	8.5	[1.26, 29.11]	[4.17, 13.69]
Wan 5B [6]	2.43	10.24	7.22	[1.76, 49.15]	[4.60, 13.05]
Wan 14B [6]	2.56	13.78	7.78	[1.75, 321.30]	[4.11, 14.11]
Gravity Adapter 5B [1]	1.28	10.27	9.74	[2.4, 26.67]	[5.98, 13.44]
Gravity Adapter 14B [1]	1.36	10.00	10.10	[2.82, 19.08]	[5.64, 13.01]

(c) Per Sample Time Scaled

Model	Mean (m/s^2)	Median (m/s^2)	Range (m/s^2)	Q1-Q3 (m/s^2)
Cosmos 2B [5]	15.08	8.69	[3.23, 128.26]	[6.57, 11.78]
Cosmos 14B [5]	15.23	11.75	[1.26, 87.06]	[6.95, 16.80]
Hunyuan [2]	17.95	10.79	[0.64, 171.55]	[6.80, 16.93]
Veo3 [4]	10.04	9.77	[5.45, 21.10]	[8.60, 11.10]
Wan 5B [6]	16.31	16.98	[2.77, 54.95]	[9.71, 20.04]
Wan 14B [6]	14.05	10.77	[1.43, 184.81]	[7.11, 15.64]
Gravity Adapter 5B [1]	10.46	10.24	[4.23, 17.17]	[8.62, 12.11]
Gravity Adapter 14B [1]	9.88	9.81	[4.88, 16.21]	[8.61, 11.07]

(d) Mean-Time Scaled with Height Adjustments

Model	MTS	Mean(m/s^2)	Median (m/s^2)	Range (m/s^2)	Q1-Q3 (m/s^2)
Cosmos 2B [5]	2.43	10.35	7.24	[1.40, 70.06]	[3.74, 12.77]
Cosmos 14B [5]	2.77	10.89	7.21	[1.81, 76.12]	[4.02, 14.21]
Hunyuan [2]	2.63	14.59	7.07	[1.55, 104.82]	[3.23, 19.71]
Veo3 [4]	2.12	9.39	8.51	[1.3, 29.12]	[4.17, 13.69]
Wan 5B [6]	2.43	10.25	7.22	[1.76, 49.29]	[4.60, 13.05]
Wan 14B [6]	2.56	13.86	7.78	[1.76, 321.56]	[4.11, 14.11]
Gravity Adapter 5B [1]	1.28	10.27	9.74	[2.4, 26.67]	[5.98, 13.44]
Gravity Adapter 14B [1]	1.36	10.00	10.13	[2.82, 19.08]	[5.64, 13.01]

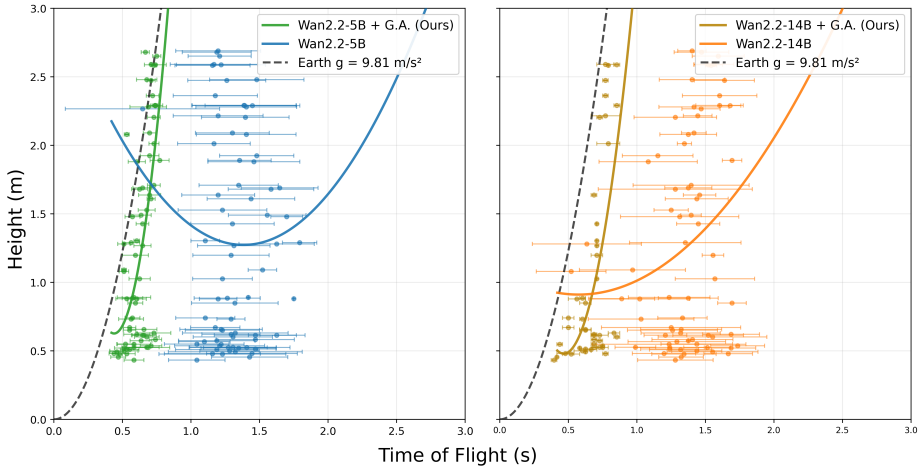
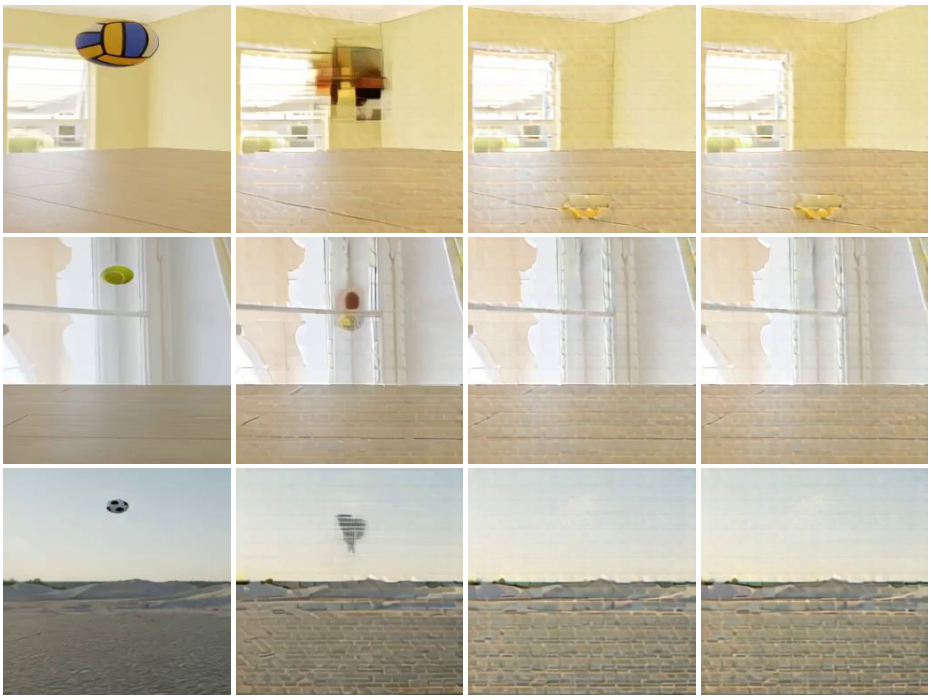
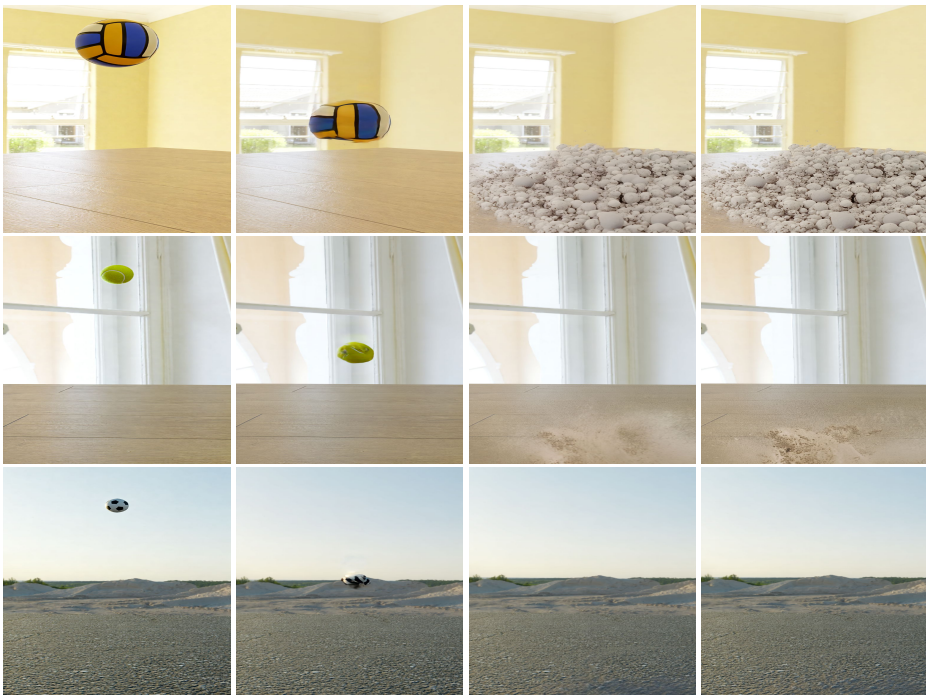


Figure 3. **Single ball falling results for gravity adapters.** We plot h versus t values for Wan 5B and Wan 14B with and without gravity adapters. The black dashed line represents terrestrial motion. We scatter plot the mean of each sample across 4 seeds and show error bars. The gravity adapters improve the base model to correct gravity while minimizing variance across seeds.

Figure 4. **Ablation on number of frames generated.** We observe that reducing the generation window to 1 second leads to poor performance. We show representative examples where either the ball remains suspended rather than falling, or the model generates an unintended extra object. Due to this, we set the generation time to 2 seconds.



Open-sora 1.2 + FT + ORO



CogVideo 1.5

time

Figure 5. **Additional models tested.** We also generate samples using OpenSora finetuned in the style of [3] and CogVideoX-1.5 [8]. However, the resulting videos exhibit severe artifacts and hallucinations, making them unsuitable for meaningful evaluation..

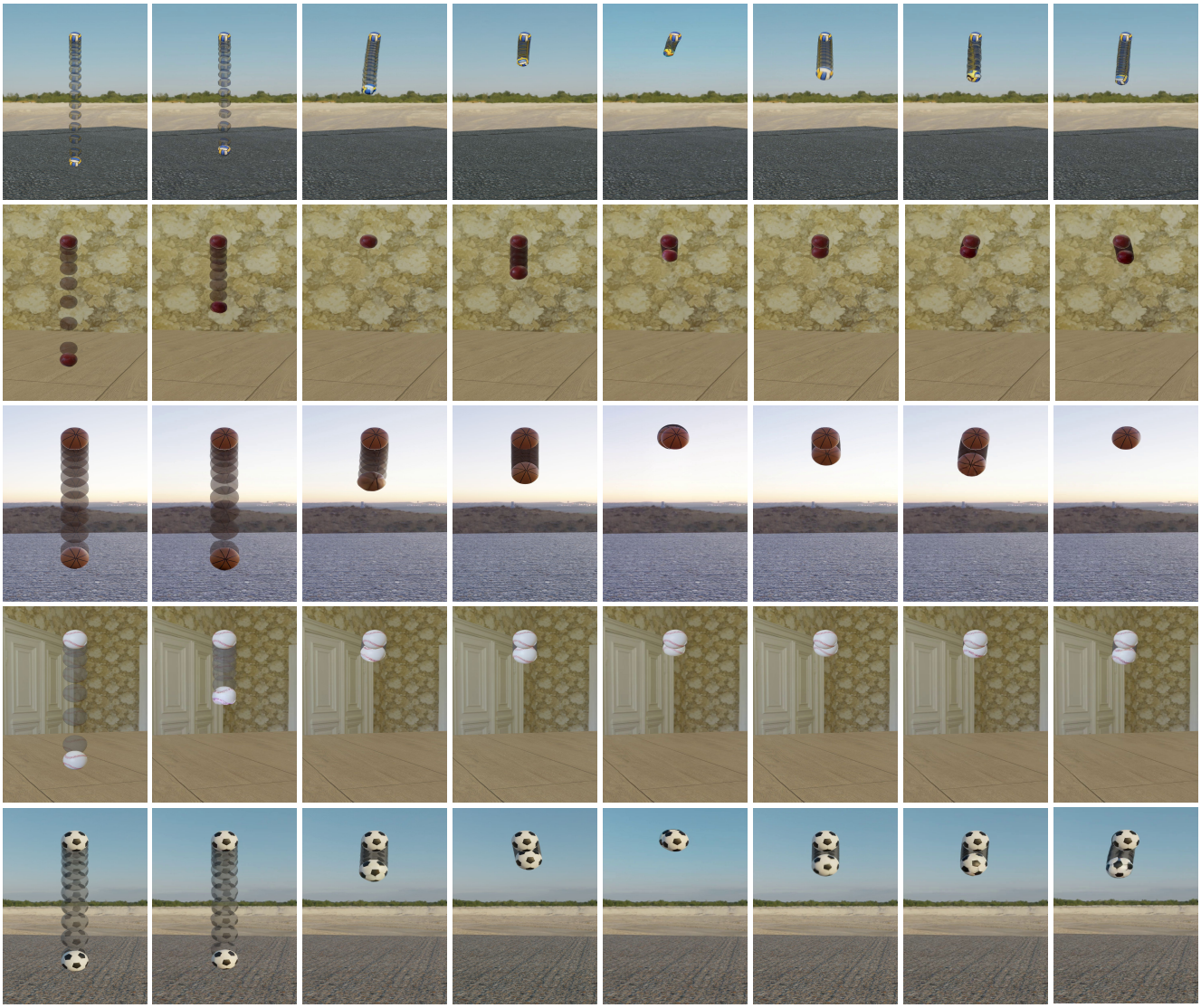


Figure 6. **Results for Single-ball drops.** Stroboscopic composites (left) visualize ball positions at equal time intervals from release. The panels show the trajectories performed by each model during the time it takes a ball falling under $9.81\ \mathrm{m/s^2}$ to reach the ground, revealing systematic under-acceleration across all models.

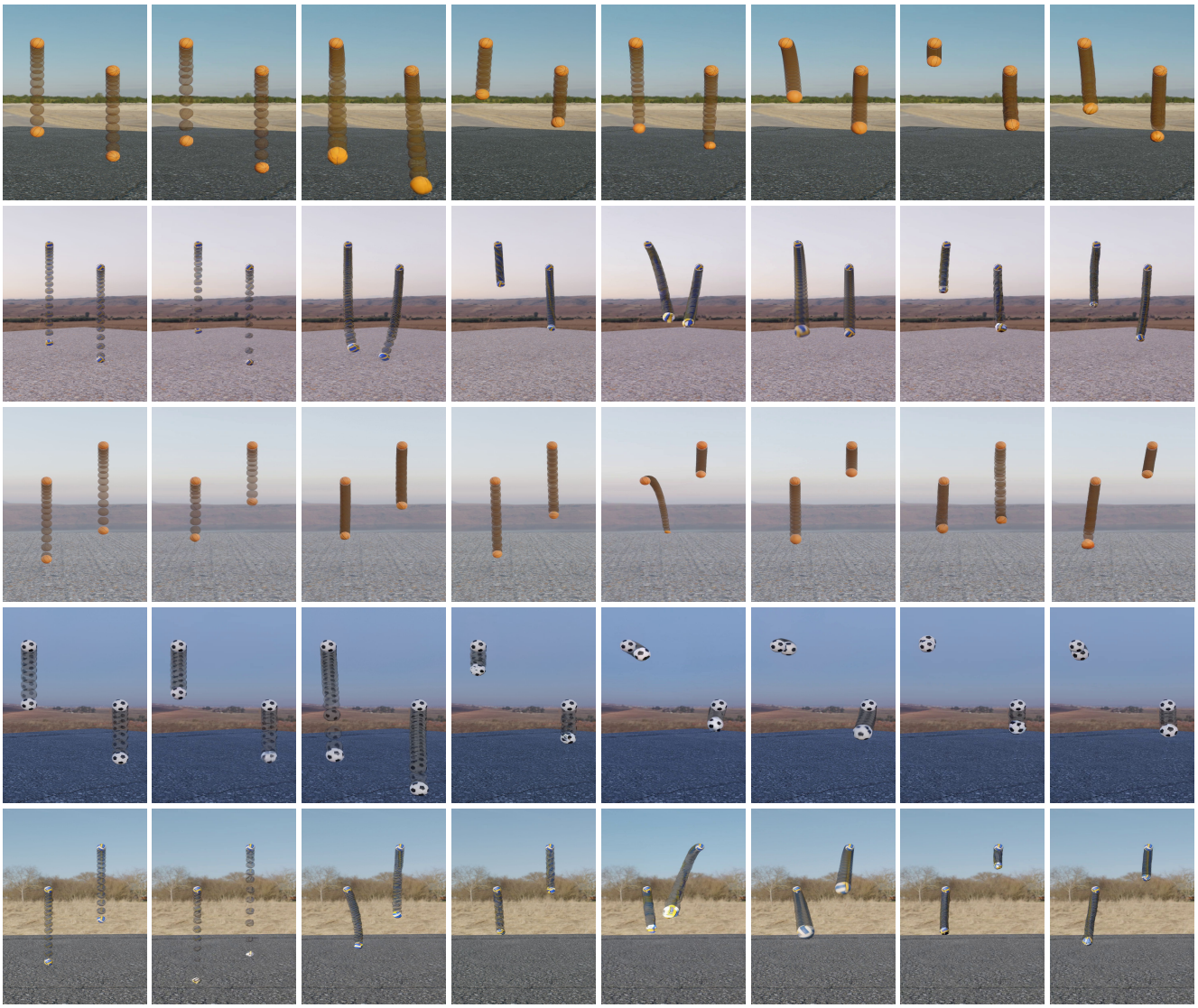


Figure 7. **Additional Results for Zero-shot generalization to double ball dropping.**

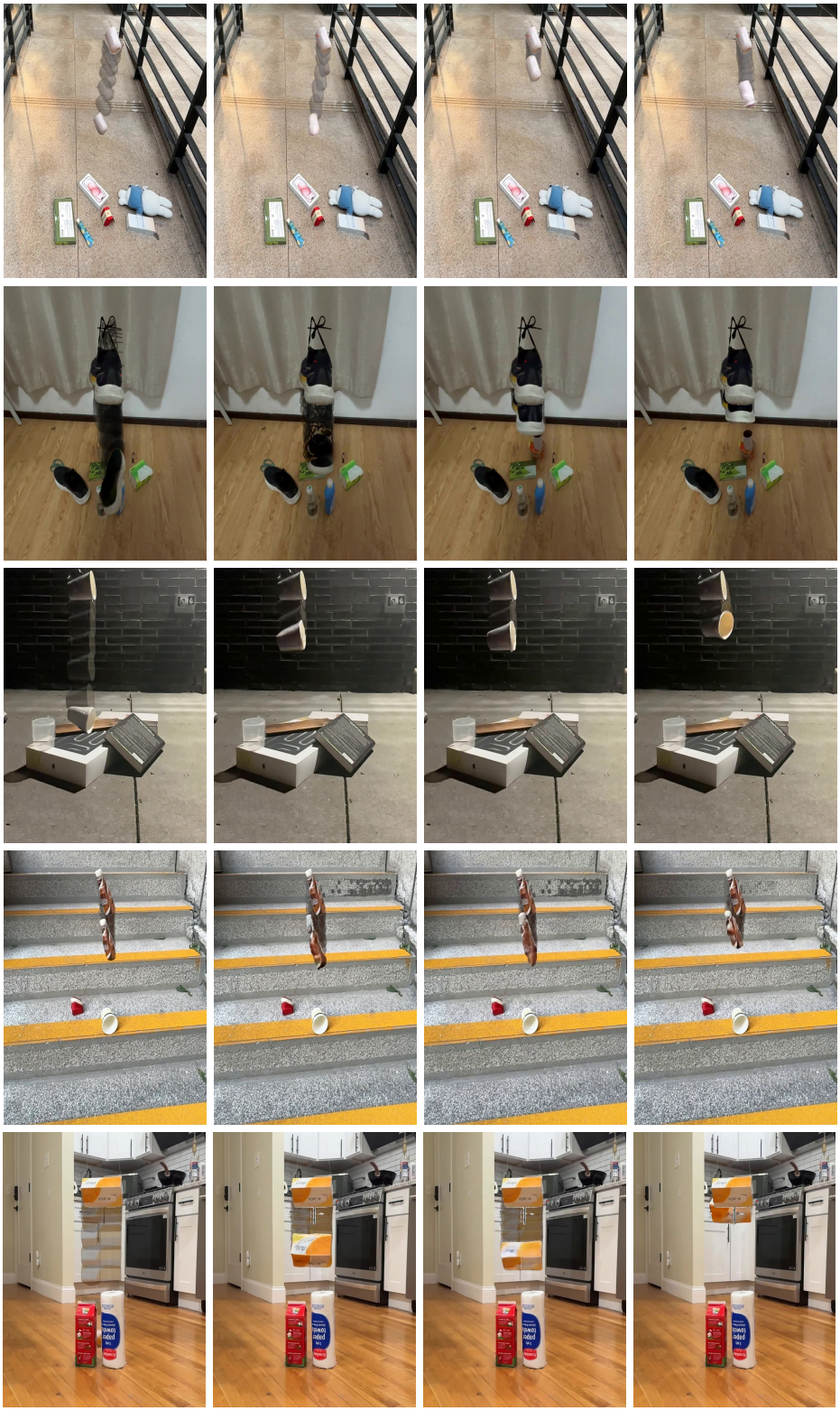


Figure 8. **Additional Results for Zero-shot generalization to real world scenes [3].**

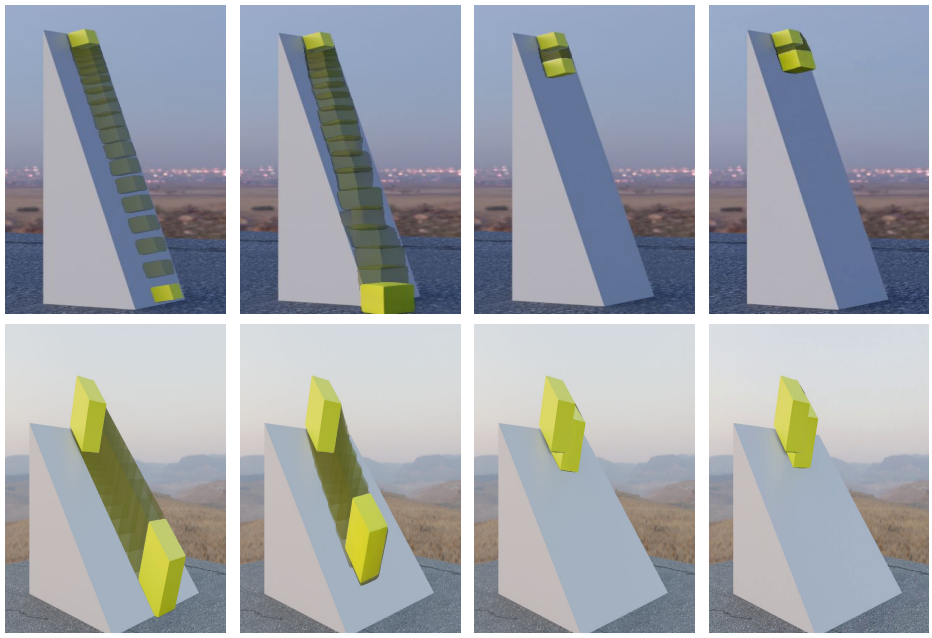


Figure 9. **Additional Results for Zero-shot generalization to inclined planes.** The gravity adapter generalizes to inclined surfaces kept at angles between 30 and 75 degrees.

References

[1] 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. 1, 3

[2] 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. 3

[3] 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. 5, 8

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

[5] 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. 3

[6] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3

[7] Angtian Wang, Haibin Huang, Zhiyuan Fang, Yiding Yang, and Chongyang Ma. ATI: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025. 1

[8] 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. 5