# SIM-∞: SIMPLE SIMULATIONS IMPROVE PHYSICS IN PHOTOREALISTIC VIDEO GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can deliberately simple, non-photorealistic simulations improve physics in photorealistic video generation? In this work, we show that a carefully designed finetuning recipe substantially improves the physical fidelity of a pretrained image-to-video model across multiple challenging physical skills, while preserving photorealism and general-domain video quality. We find that visually abstract, minimally textured simulations paired with detailed captions effectively teach targeted physical skills. Meanwhile, co-training with randomly sampled real videos preserves the model's photorealism. Our experiments investigate three challenging physical skills: object permanence under occlusion, rigid-body collision dynamics, and deformable-body wave propagation. On a benchmark with independently constructed photorealistic physical scenes, a single model trained jointly on all three skills significantly outperforms a real-video-only baseline across all physics types. Physics-IQ-Verified increases from 47.98 to 49.80, while VBench++ scores remain closely matched. Ablations examine how the simulation-to-real-video ratio and caption design affect skill transfer and appearance retention. The method requires neither expensive photorealistic simulation rendering nor a simulator at inference time, offering a scalable way to expand the physical capabilities of video world models.

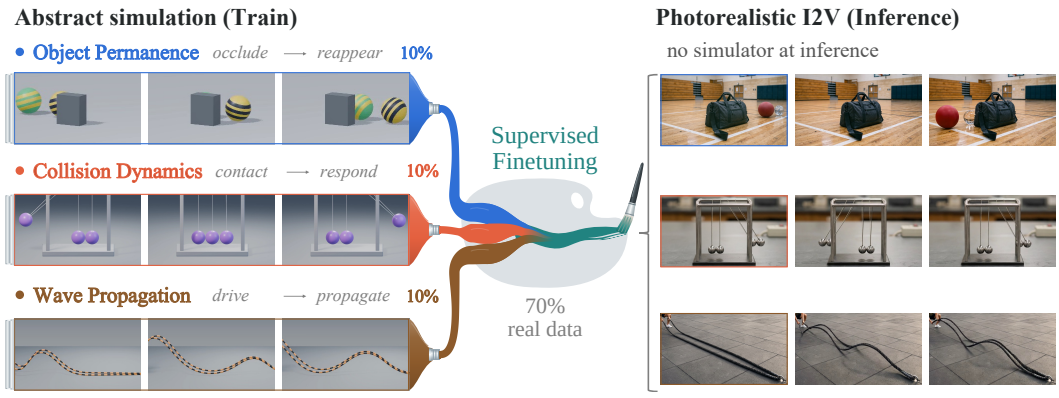


Figure 1: **Sim-∞: Simulation-to-photorealistic skill transfer.** The model is trained on abstract simulations (left) and can generate photorealistic scenes (right) with the target physical skills.

## 1 INTRODUCTION

Video diffusion models (ByteDance, 2026; NVIDIA, 2026; MiniMaxAI, 2026; Google DeepMind, 2026a) have demonstrated strong capability in generating photorealistic videos with convincing visual appearance, yet even the best models still struggle with physical behavior. An object may fail to reappear or unnaturally morph after occlusion, a ball may fail to rebound promptly after an elastic collision, or a deformable rope may break apart during large-amplitude waves. These failures matter when video models are used as world simulators: visual realism alone does not ensure that a scene evolves according to consistent physical behavior. Improving physical consistency while retaining



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

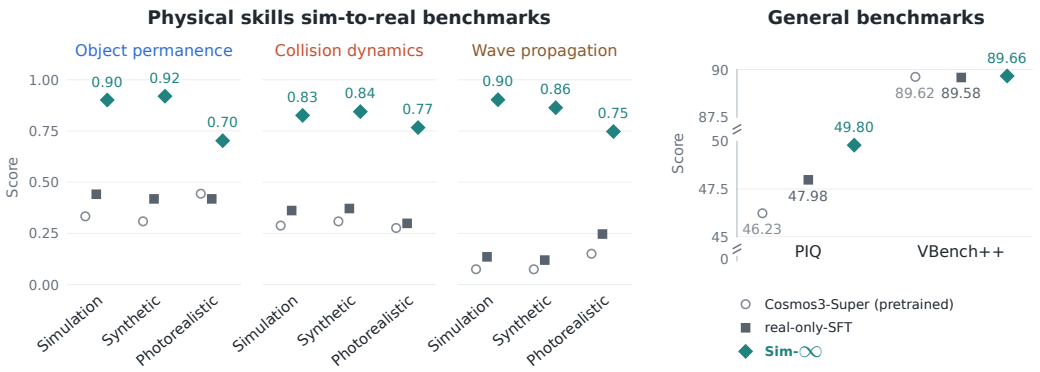


Figure 2: **Joint skill acquisition, sim-to-photorealistic transfer, and general benchmark performance.** Left: Sim-$\infty$ improves over the pretrained Cosmos3-Super (NVIDIA, 2026) and real-only-SFT across all nine skill/domain combinations. Right: Sim-$\infty$ improves over the pretrained Cosmos3-Super (NVIDIA, 2026) and real-only-SFT on Physics-IQ-Verified (PIQ) (Motamed et al., 2026; Rädsch et al., 2026) and achieves comparable scores on VBench++ (Huang et al., 2026).

the visual capabilities of video diffusion models is therefore a central challenge for establishing video diffusion models as reliable world simulators.

Developing a reliable world simulator inherently requires high-quality supervision that clearly demonstrates physical laws. However, obtaining data that cleanly isolates intended physical behaviors from real-world complexity is profoundly challenging. Real-world videos inextricably entangle appearance, lighting, camera viewpoint, motion, and complex interactions, precluding controlled experiments on isolated physical skills. While simulation environments allow for precise control over scene configurations to isolate target behaviors, recent simulation-based post-training approaches (Li et al., 2025a; Zhao et al., 2025; Yuan et al., 2026; Romero et al., 2026) predominantly rely on computationally expensive photorealistic rendering, target only isolated skills, or demand additional inference-time guidance. Since pretrained video models already possess rich visual priors, a fundamental question remains: *Can deliberately simple, non-photorealistic simulations effectively instill physical skills that transfer to photorealistic video generation?*

To address this, we introduce *Sim-*$\infty$, a post-training recipe that leverages simple, physics-targeted simulations mixed with real videos not curated for the target skills to enhance physics in photorealistic video generation. We target three challenging physical skills: object permanence under occlusion (OP), rigid-body collision dynamics (CD), and wave propagation along deformable ropes (WP). Our simulated scenes intentionally utilize basic geometric shapes and non-photorealistic rendering. Each simulation is paired with a highly descriptive caption detailing the initial state, object kinematics, interactions, and final outcomes. We then fine-tune Cosmos3-Super (NVIDIA, 2026) using these simulation–caption pairs mixed with 70% real-world videos not curated for the target skills. This methodology supports both single-skill specialists and a unified model—denoted Sim-$\infty$—without requiring a simulator or additional guidance during inference.

We evaluate skill acquisition in simulated, synthetic, and photorealistic domains, as well as broader physical common sense via Physics-IQ-Verified (Motamed et al., 2026; Rädsch et al., 2026) and general video quality via VBench++ (Huang et al., 2026). Prior to post-training, Cosmos3-Super (NVIDIA, 2026) lags behind the open-source MiniMax-H3 (MiniMaxAI, 2026) by margins of 0.27, 0.26, and 0.58 on photorealistic OP, CD, and WP, respectively, and by 3.9 points on Physics-IQ-Verified. Following joint fine-tuning on all three skills, Sim-$\infty$ effectively closes these gaps: it performs on par with MiniMax-H3 on OP, WP, and Physics-IQ-Verified ($-0.01$, $+0.02$, and $-0.35$), while establishing a 0.23 lead in CD. Furthermore, Sim-$\infty$ surpasses the proprietary Gemini Omni Flash (Google DeepMind, 2026a) across all three targeted skills and Physics-IQ-Verified, all while maintaining its baseline VBench++ general video quality (Table 2). Our analyses (Sec. 3) detail how Sim-$\infty$ successfully transfers these skills from abstract supervision to photorealistic generation, highlighting the critical role of the simulation-to-real mixture ratio and caption design in balancing task performance and appearance retention.

Our findings demonstrate that simple simulations, when mixed with real videos, provide scalable and effective supervision for instilling complex physical skills into photorealistic video generation models. Our core contributions are:

- We introduce Sim-$\infty$, a fine-tuning recipe that combines 30% simple, physics-targeted simulations with 70% real-video mixing to significantly improve physics in photorealistic video generation.
- We establish a scalable simulation data pipeline to construct a 297k-clip physics training dataset across three core skills, and design a comprehensive evaluation benchmark across simulation, synthetic, and photorealistic domains with task-specific VLM metrics.
- We demonstrate joint acquisition of three physical skills: compared with real-video-only fine-tuning, a single model improves all nine skill-domain combinations and Physics-IQ-Verified scores while maintaining similar VBench++ scores. Single-skill training reveals selective transfer, improving some untrained skills while degrading others.

## 2 SIM-$\infty$: SIM-TO-PHOTOREALISTIC SKILL INJECTION

**Motivation.** We manually review outputs from MiniMax-H3 (MiniMaxAI, 2026), Seedance 2.0 (ByteDance, 2026), Cosmos3-Super (NVIDIA, 2026), and Gemini-Omni (Google DeepMind, 2026a) on Physics-IQ-Verified (Motamed et al., 2026; Rädsch et al., 2026), PAI-Bench (Zhou et al., 2026), and additional generation cases to identify recurring physics issues, visualized in Appendix C. We systematically analyze the observed failures using first-principles reasoning and group those sharing the same underlying physics into categories, including object permanence, collision dynamics, wave propagation and direction control, and fluid conservation. Our frequency analysis identifies object permanence and collision dynamics as the two most common failure categories, motivating their selection as case studies. We additionally include wave propagation as an exploratory case study to extend our investigation to challenging dynamics in deformable media. For these three selected skills, we define target judgement criteria that specify the expected physical behavior and guide the design of minimal simulation scenes. Our study asks two questions:

- Can simulation data inject scalable targeted physical skills into a pretrained I2V model?
- Can these skills transfer from abstract simulation scenes to photorealistic scenes while preserving general generation capabilities?

The training scenes are deliberately non-photorealistic, as illustrated in Fig. 1.

**Overall pipeline.** We design simulation scenes for each target skill, prepare training captions, and post-train the model with the simulation data. We then construct simulation, synthetic, and photorealistic evaluation scenes sharing the same target physics. Finally, we design task-specific metrics to assess the injected skills.

| Simulation scene design | → | Caption preparation | → | Model post-training | → | Evaluation scene design | → | Evaluation metric design | → | Model evaluation |
|---|---|---|---|---|---|---|---|---|---|---|
| (Sec. 2.1) | | (Sec. 2.2) | | (Sec. 2.2) | | (Sec. 2.3) | | (Sec. 3.1) | | (Secs. 3.2–3.4) |

### 2.1 PHYSICS-TARGETED SIMULATION DESIGN

Before starting each simulation design, we first formulate the judgement criterion for the target physics. Given the observed failure cases, we disentangle the underlying physics and vary scene factors (*e.g.*, object counts, shapes, speeds and colors) to construct the simulation scenes. We construct the simulation scenes in Blender (Blender, 2026) and render them with EEVEE at $1280 \times 720$ resolution and 24 FPS. Scene-specific settings are provided in Appendix A.1.

#### 2.1.1 OBJECT PERMANENCE (OP) UNDER OCCLUSION

**Judgement Criterion (Fig. 3).** To address disappearance, duplication, and identity changes during occlusion, moving objects should reappear with the same count and identity and continue with coherent motion while stationary objects and the scene remain stable.

**Data Design (Fig. 10 in Appendix A.1).** We construct 98,280 occlusion configurations in which one to five rolling spheres pass behind one to three stationary occluders and reappear. We vary

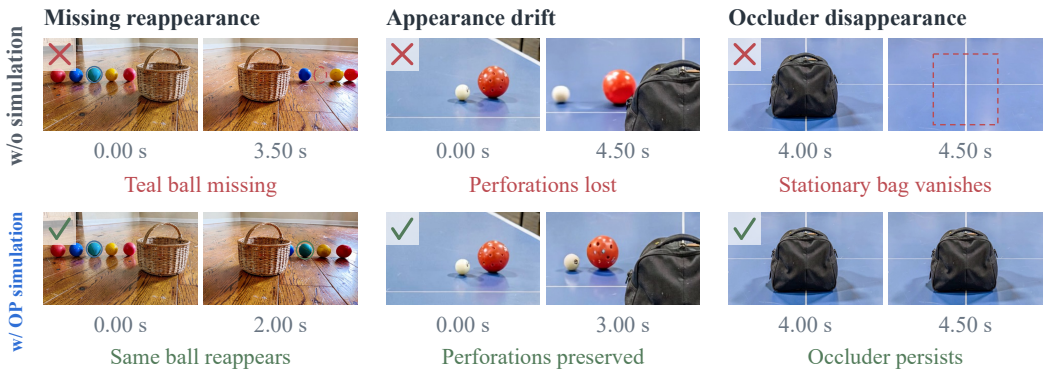


Figure 3: **Object Permanence negative and positive cases.**

object count and spacing, surface markings, speed, occluder shape and placement, and three fixed camera views; sphere colors, a shared radius, initial visibility, and exact spacing vary within each grid cell. Occluders produce either complete hiding or heavy partial occlusion, making persistence through different visibility conditions observable. Cameras remain stationary, and objects maintain sufficient clearance from occluders throughout their trajectories to prevent contact, allowing us to study continuity through occlusion without camera motion or object–occluder collisions.

### 2.1.2 Rigid-body collision dynamics (CD)

**Judgement Criterion (Fig. 4).** To address delayed or sticky rebounds, incorrect motion transfer, and broken collision cascades, contact should produce an immediate, causally coherent response while preserving the participating objects.

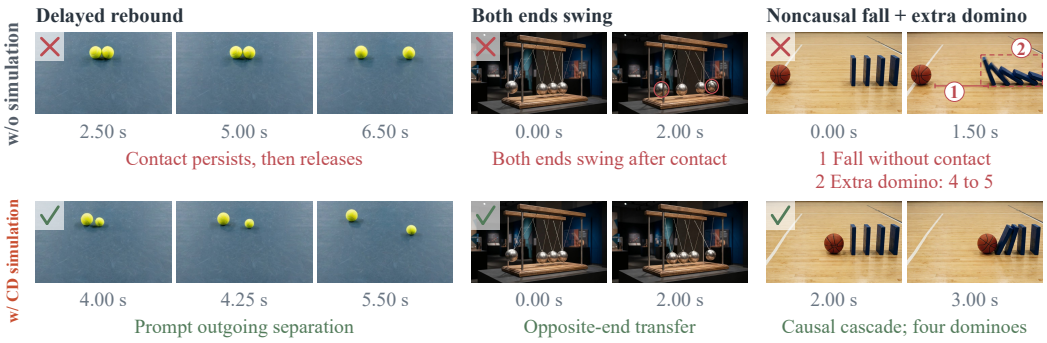


Figure 4: **Collision Dynamics negative and positive cases.**

**Data Design (Fig. 11 in Appendix A.1).** We construct 99,300 clips covering ball impacts, Newton's cradle, and ball-induced domino toppling (69,300, 15,000, and 15,000 clips, respectively). Ball impacts involve other balls or fixed obstacles. We vary object count, ball size and mass, initial motion, and incidence angle, rendering each physical rollout from three fixed camera views. Scenes include visible approach, contact, and response phases so that delayed rebounds, incorrect motion transfer, and interrupted cascades can be distinguished. We reject interpenetration, spurious motion, and unreasonable energy gain.

### 2.1.3 Deformable-body dynamics: wave propagation (WP)

**Judgement Criterion (Fig. 5).** To address jerky or noncausal rope motion and breaks in rope continuity, controller motion should initiate smooth, sustained traveling waves along an intact rope.

**Data Design (Fig. 12 in Appendix A.1).** We construct 99,960 clips from 14,280 physical rollouts spanning seven motion families: single-leg vertical and lateral waves, and two-leg alternating, synchronous, inward-circular, outward-circular, and lateral waves. Across these scenes, pe-

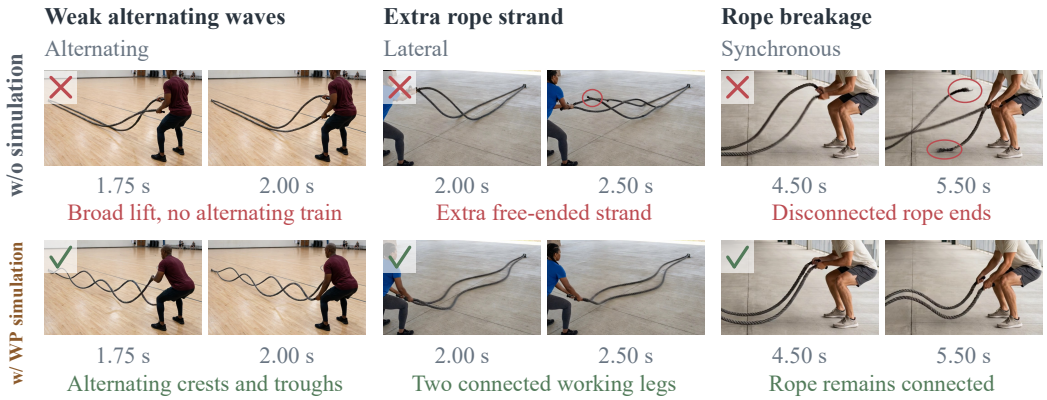


Figure 5: **Wave Propagation negative and positive cases.**

riodic controller motion produces smooth, sustained waves that travel along the rope away from the driven ends. We vary controller frequency and motion amplitude together within standard- and high-intensity groups, producing waves with different amplitudes and wavelengths. We also vary controller and rope colors and rope surface markings, rendering each rollout from seven fixed perspective views.

### 2.2 Caption preparation and post-training recipe

**Caption preparation.** We use simulation metadata to prepare training captions describing the initial scene and participating objects, the target physics and motion, and the outcome and final scene state, using Qwen3.8-27B (Qwen Team, 2026). The captions explicitly state that the video is a `simulation`. Section 3.4 ablates the explicit simulation wording and the outcome description.

**Post-training recipe.** We fine-tune the pretrained I2V model, Cosmos3-Super (NVIDIA, 2026), with 30% simulation video–caption pairs and 70% real videos not curated for the target skills for around two simulation data exposure epochs with a batch size of 128. For the specialist models (OP-SFT, CD-SFT, and WP-SFT), the simulation portion comes from one target skill; for Sim-$\infty$, it is divided equally among the three skills (10% each of object permanence, collision dynamics, and wave propagation). We ablate the simulation mixture percentage in Section 3.4, and the training epochs are detailed in Appendix A.2.

### 2.3 Evaluation scene construction

To assess sim-to-photorealistic transfer, we construct evaluation scenes in three domains (Table 1; Fig. 6). **Simulation** scenes use the same simulation pipeline as the training data, with unique object placements and parameter configurations. **Synthetic** scenes are image-to-image translations of these frames using FLUX-Klein-4B (Black Forest Labs, 2026), HiDream-O1-Image (HiDream-ai, 2026), or Nano Banana 2 (Google DeepMind, 2026b). They preserve task-relevant objects and

Table 1: Evaluation domains.

| Domain | Task-relevant scene content | Image appearance |
|---|---|---|
| Simulation | Same scene types; new parameters | Simulator rendering |
| Synthetic | Preserved from simulation | I2I-translated |
| Photorealistic | New objects and scenes | Photorealistic |

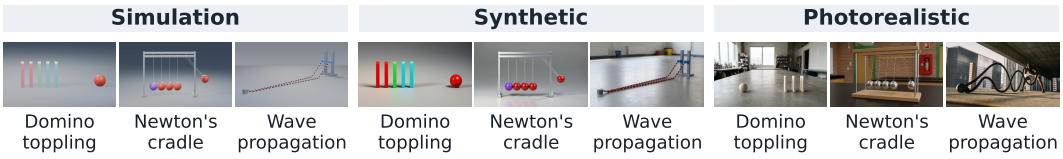


Figure 6: **Evaluation examples.** Simulation and synthetic frames are paired; photorealistic scenes are independently constructed. Appendix A.3 provides additional examples.

layouts while varying textures, lighting, and backgrounds. **Photorealistic** scenes are independently constructed for the same physical activities with new objects and environments; this set also includes real photographs.

# 3 Experiments and Analysis

## 3.1 Experimental setup and metrics

**Models.** We use the Cosmos3-Super (Pretrained) (NVIDIA, 2026) and a version fine-tuned only on real videos (real-only-SFT) as the base comparisons with the latter serving as our primary reference. To assess individual and joint skill acquisition, we evaluate three specialist models (OP-SFT, CD-SFT, and WP-SFT) and a joint model Sim-$\infty$, trained using the recipe in Sec. 2.2. We additionally compare Sim-$\infty$ with Gemini Omni Flash (Google DeepMind, 2026a) and MiniMax-H3 (MiniMaxAI, 2026) on photorealistic physics tasks and the general benchmarks.

**Evaluation protocol.** We evaluate performance on two general benchmarks, VBench++ (Huang et al., 2026) and Physics-IQ-Verified (PIQ) (Motamed et al., 2026; Rädsch et al., 2026), and three specialized benchmarks, object permanence, collision dynamics, and wave propagation, across the simulation, synthetic, and photorealistic domains with 3 generation seeds at 832×480 resolution and 24 FPS. For OP, we construct 100 scenes per domain, resulting in 300 scenes and 900 videos per model. For CD, we construct 100 scenes for each of the three interaction families (ball impacts, Newton's cradle, and domino toppling), resulting in 900 scenes and 2,700 videos per model. For WP, we construct 210 scenes for simulation and synthetic evaluation and 178 scenes for photorealistic evaluation, resulting in 598 scenes and 1,794 videos per model. For PIQ, to make it better aligned with the I2V settings, we rewrite each prompt as an I2V instruction conditioned on the first frame; details are in Appendix A.4. VBench++ evaluates 355 non-camera cases, with more details in Appendix A.5.

**Task-specific metrics.** We design task-specific GPT judges (OpenAI, 2026a;b) for each task to inspect generated videos and assign a holistic completion score in $[0, 1]$, following the physical criteria in Section 2.1; details are in Appendix A.6. To develop the judging guidance, we manually score 10 videos from each model with model identities hidden to calibrate score ordering and differences. The automatic and human evaluation alignment analysis is reported in Appendix B.1. Details about the Physics-IQ-Verified (Motamed et al., 2026; Rädsch et al., 2026) and VBench++ (Huang et al., 2026) scoring protocols are reported in Appendices A.4 and A.5.

## 3.2 Skill acquisition and cross-domain transfer

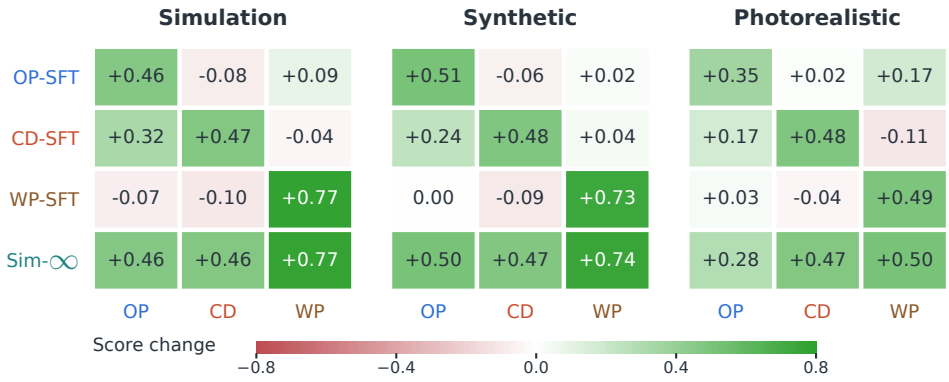


Figure 7: **Skill specificity and joint acquisition.** Each cell gives the score difference from real-only-SFT on the same task and domain. Columns denote evaluation tasks; rows denote training variants. Positive values indicate higher completion scores. The common color scale preserves the magnitude of changes across panels.

**Specialists show cross-domain gains and selective cross-task transfer.** Fig. 2 summarizes the acquisition of each target skill and Fig. 7 shows the performance on both trained and untrained skills. Relative to real-only-SFT, the matching specialists improve simulation scores by +0.46 (OP), +0.47

(CD), and +0.77 (WP). The corresponding photorealistic gains remain positive at +0.35, +0.48, and +0.49. These results support sim-to-photorealistic transfer for the same target skills. However, the transfer between skills is selective for the specialist model post-training. CD-SFT improves OP by +0.32, +0.24, and +0.17 across simulation, synthetic, and photorealistic scenes while OP-SFT improves photorealistic WP by +0.17, whereas CD-SFT reduces it by 0.11. CD and WP are negatively correlated in simulation, synthetic, and photorealistic scenes.

**Scale-up: Joint acquisition of multiple physical skills.** We train Sim-$\infty$ with supervision for all three physical skills. Sim-$\infty$ improves over real-only-SFT in all nine task/domain combinations (Figure 7) with gains comparable to those of the specialists in all domains. Across eight of the nine settings, the joint score differs from the matching specialist by at most 0.02; photorealistic OP is the exception, with a 0.07 lower score. This provides evidence that one model can acquire the three skills together, indicating a positive sign for scaling up to more skills. The statistical reliability of performance gains is reported in Appendix B.3.

### 3.3 Photorealistic benchmarks comparison

Table 2 reports results on three photorealistic physics tasks and two general benchmarks. Sim-$\infty$ improves over Pretrained on all five. Furthermore, trained on abstract simulation data, Sim-$\infty$ outperforms proprietary Gemini Omni Flash (Google DeepMind, 2026a) on all three physics tasks and achieves scores comparable to or better than those of open-source MiniMax-H3 (MiniMaxAI, 2026) with qualitative examples in Fig. 8. On PIQ (Motamed et al., 2026; Rädsch et al., 2026), Sim-$\infty$ achieves relative gains of +7.71% and +3.78% over Pretrained and real-only-SFT, respectively. Its VBench++ (Huang et al., 2026) score exceeds those of Pretrained and real-only-SFT by 0.04 and 0.08, respectively.

Table 2: **Photorealistic physics and general benchmarks comparison.**

| | Photorealistic physics | | | General benchmarks | |
|---|---|---|---|---|---|
| Model | OP | CD | WP | PIQ | VBench++ |
| MiniMax-H3 *(open-source)* | **0.71** | 0.54 | 0.73 | **50.15** | 88.35 |
| Gemini Omni Flash *(proprietary)* | 0.42 | 0.39 | 0.48 | 45.67 | **90.01** |
| Cosmos3-Super (Pretrained) | 0.44 | 0.28 | 0.15 | 46.23 | 89.62 |
| real-only-SFT | 0.42 | 0.30 | 0.25 | 47.98 | 89.58 |
| **Sim-$\infty$ (ours)** | 0.70 | **0.77** | **0.75** | 49.80 | 89.66 |
| (vs. Pretrained) | (+58.34%) | (+177.27%) | (+395.77%) | (+7.71%) | (+0.05%) |

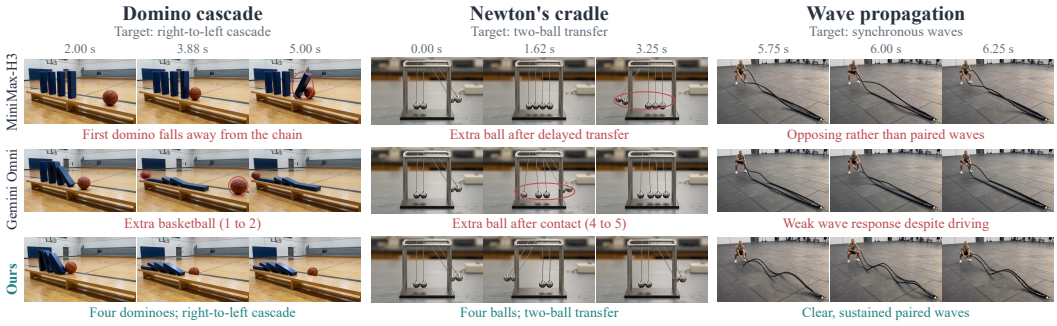


Figure 8: **Qualitative comparison.** Ours denotes Sim-$\infty$. Red annotations highlight wrong-direction toppling, duplicated objects, and mismatched or weak rope waves.

**Failures under the holistic score.** Similar holistic completion scores can hide distinct failure profiles in Table 2. In the photorealistic domain, we group related verdict diagnostics and rank categories by their mean per-video incidence across Pretrained, MiniMax-H3 (MiniMaxAI, 2026), and Gemini Omni (Google DeepMind, 2026a), weighting these models equally. Table 3 reports the top three failures per task and Sim-$\infty$ reduces all nine diagnosed failure rates relative to Pretrained or real-only-SFT. Interestingly, Sim-$\infty$ and MiniMax-H3 (MiniMaxAI, 2026) have similar overall scores (0.75 vs. 0.73), but motion-family/phase/direction errors differ markedly (46.3% vs. 84.3%). Thus, generating coherent traveling waves and following the requested motion remain distinct challenges. However, Sim-$\infty$ exhibits no detected motion-family/phase/direction errors in simulation

(0/630), compared with 46.3% (247/534) in the photorealistic domain. This gap may reflect the mismatch between controller-driven ropes in simulation and human-driven ropes in the photorealistic domain, as well as ambiguity in prompt descriptions of different motion families. Future work could explore simulation data that better captures human actuation and caption ablations to assess whether more explicit motion-family descriptions improve control. Despite this gap, Sim-$\infty$ with our simulation data substantially improves both motion control and wave propagation: relative to the pretrained model, motion-control errors decrease by 33.3 percentage points (79.6% to 46.3%), while wave formation/propagation failures fall from 86.0% to 4.7%. Appendix B.2 provides temporal examples.

Table 3: **Failure rates of top-3 issue groups per task** on photorealistic videos (%, lower is better).

| Task | Issue | Pretrained | real-only-SFT | Sim-$\infty$ | MiniMax-H3 | Gemini Omni |
|---|---|---|---|---|---|---|
| OP | Unnatural motion/rolling | 51.0 | 36.0 | **14.0** | 40.7 | 39.7 |
| | Object/scene inconsistency | 21.7 | 41.7 | **15.0** | 15.7 | 58.0 |
| | Incorrect occlusion | 47.7 | 39.7 | 12.0 | **6.0** | 19.0 |
| CD | Incorrect collision response | 88.9 | 82.9 | **21.7** | 55.8 | 59.1 |
| | Object/structural inconsistency | 34.0 | 41.8 | 13.2 | **9.0** | 49.1 |
| | Unnatural/discontinuous motion | 35.4 | 33.4 | **6.2** | 9.3 | 38.4 |
| WP | Motion-family/phase/direction error | 79.6 | 80.7 | **46.3** | 84.3 | 83.7 |
| | Material frequency mismatch | 81.1 | 88.6 | **3.9** | 8.6 | 61.4 |
| | Wave formation/propagation failure | 86.0 | 68.0 | **4.7** | 6.9 | 20.2 |

*Note: Evaluated videos per model: OP 300; CD 900; WP 534. Each video counts once per category.*

### 3.4 Ablations

**Simulation data mixture ratios.** Table 4 compares 30% and 100% simulation data mixtures for Collision Dynamics training. At matched simulation exposure, CD scores differ by at most 0.02 and PIQ scores by 0.01, but VBench++ drops by 0.34 points (89.74 to 89.40) with 100% simulation data. With 100% simulation data, increasing exposure from 1× to 3.3× yields no further in-domain gains and lowers PIQ from 48.13 to 46.82, while also reducing VBench++ dynamic degree (Table 10).

Table 4: Simulation data mixture ablation on Collision Dynamics.

| Mixture ratio | Sim. exposure | Train steps | Sim. | Syn. | Photo. | PIQ | VBench++ |
|---|---|---|---|---|---|---|---|
| Pretrained | – | – | 0.29 | 0.31 | 0.28 | 46.23 | 89.62 |
| 30% | 1× | 2,500 | 0.83 | 0.85 | 0.78 | 48.14 | 89.74 |
| 100% | 1× | 750 | 0.85 | 0.84 | 0.78 | 48.13 | 89.40 |
| 100% | 3.3× | 2,500 | 0.82 | 0.83 | 0.79 | 46.82 | 89.42 |

The qualitative examples in Fig. 9 further highlight the importance of mixing real-world data into training to preserve photorealistic appearance. With 100% simulation data, objects visible in the first frame largely retain their initial appearance, whereas newly entering objects appear more simulation-like. In the bottom row, simulation-only training produces smooth spheres with faint shadows, while the 30% simulation mixture preserves visible tennis ball seams and more distinct shadows.

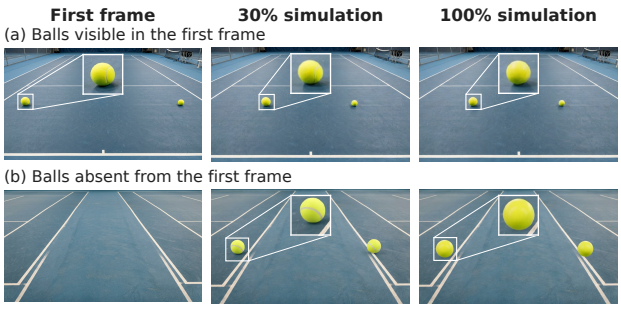


Figure 9: **Appearance of existing and newly entering objects.** Both models are compared at the same simulation exposure. In the bottom row, 100% simulation produces smooth spheres with weaker shadows; 30% retains visible seams and more distinct shadows.

**Captions and prompts.** Table 5 compares two ablations with the standard OP-SFT configuration. In the *Without "simulation"* variant, we remove the word `simulation` from the training captions; the simulation videos remain in the training data. This lowers the photorealistic score from 0.77 to 0.75.

The *No outcome guidance* variant uses the same OP-SFT checkpoint but omits outcome descriptions from inference prompts. Scores decrease across all three domains but remain above those of the Pretrained model evaluated with outcome-explicit prompts. These results suggest that simulation training supports skill acquisition, while outcome-explicit prompts provide further gains.

Table 5: Caption and prompt ablations.

| Variant | Sim. | Syn. | Photo. |
|---|---|---|---|
| Pretrained | 0.33 | 0.31 | 0.44 |
| OP-SFT | 0.90 | 0.92 | 0.77 |
| Without "simulation" (training captions) | 0.88 | 0.92 | 0.75 |
| No outcome guidance (inference prompt) | 0.59 | 0.60 | 0.53 |

Shortened examples for the same scene illustrate the three variants: *OP-SFT*: "A simulation video shows five spheres rolling left behind three occluders. Each reappears after occlusion; three remain visible in the final frame." *Without "simulation"* (training caption): "A video shows five spheres rolling left behind three occluders. Each reappears after occlusion; three remain visible in the final frame." *No outcome guidance* (inference prompt): "A video shows five spheres rolling left behind three occluders."

## 4 Related Work

**Simulation-Guided Video Generation.** Recent efforts leverage physical simulation and synthetic data to improve video fidelity and controllability. One direction focuses on *inference-time guidance*: PhysGen (Liu et al., 2024) and PSIVG (Foo et al., 2026) use training-free physical simulators to guide trajectories and refine textures, while NewtonGen (Yuan et al., 2026) models continuous neural ODE dynamics to steer synthesis via optical flow. A second direction applies *explicit physics conditioning*: Force Prompting (Gillman et al., 2025) models responses to applied forces, whereas Goal Force (Gillman et al., 2026) generates interactions that achieve a target force. PhysCtrl (Wang et al., 2025a) guides synthesis with 3D trajectories generated by a simulation-trained model. KineMask (Romero et al., 2026) trains ControlNets on simulated velocity masks, PhyCo (Narayanan et al., 2026) combines physical-property-map conditioning with VLM-guided rewards, and WISA (Wang et al., 2025b) incorporates dedicated physical modules. Beyond explicit control, simulation data provides supervision for studying dynamics generalization (Kang et al., 2025), enhancing consistency through mixed synthetic-real fine-tuning (Zhao et al., 2025), or acquiring isolated skills like simulated freefall (Li et al., 2025a). In contrast, our framework requires neither inference-time simulators nor explicit force conditioning; fine-tuning on abstract, non-photorealistic simulations mixed with real video enables targeted physical skills to transfer directly to photorealistic image-to-video generation.

**Physical Evaluation and Failure Diagnosis.** Physical evaluation spans reference-based baselines and VLM judges. Physics-IQ-Verified (Motamed et al., 2026; Rädsch et al., 2026) compares generated dynamics against ground-truth physical experiments, whereas VideoPhy and VideoPhy-2 (Bansal et al., 2025; 2026) assess physical commonsense alongside semantic adherence. Other benchmarks analyze physical law violations using human-supervised judges (Li et al., 2025b; Lin et al., 2026) or hierarchical VLM event checks (Meng et al., 2025). To evaluate our diverse scenarios, we design task-specific rubrics evaluated by off-the-shelf VLMs, enabling automated physical inspection without fine-tuning dedicated judges.

## 5 Conclusion

Sim-$\infty$ introduces a scalable supervised fine-tuning recipe that utilizes deliberately simple, non-photorealistic simulations to successfully inject complex physical skills—specifically object permanence, collision dynamics, and wave propagation—into pretrained image-to-video models. By pairing these visually abstract, physics-targeted scenes with descriptive captions and mixing them with real-world videos not curated for the target skills, we enable a single model to jointly acquire these skills without requiring inference-time simulators or expensive photorealistic renderings. This joint training significantly outperforms real-video-only fine-tuning across simulation, synthetic, and photorealistic domains, improving physics-specific benchmarks like Physics-IQ-Verified while preserving general video generation quality on VBench++. Ultimately, these findings establish that deliberately abstract simulation data offers a practical foundation for enhancing the physical fidelity of generative world models, paving the way for future extensions into dynamic camera movements and robotic tasks.

### AI USE STATEMENT

We used generative AI tools to prepare synthetic evaluation scenes, generate captions, and assist with operating simulation tools and evaluation workflows. We did not use these tools to propose or refine hypotheses, design or critique research methods or experiments, or interpret results. Developing theoretical models or conceptual frameworks and formulating or proving mathematical claims are outside the scope of this work. We manually reviewed all AI-assisted work, including synthetic scenes and captions, and take responsibility for the final content, claims, and artifacts.

### REPRODUCIBILITY STATEMENT

We describe the simulation setup in Appendix A.1, evaluation scenes in Appendix A.3, and evaluation procedures in Appendices A.4, A.5, and A.6.

## REFERENCES

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Black Forest Labs. FLUX.2-klein-4B. `https://huggingface.co/black-forest-labs/FLUX.2-klein-4B`, 2026.

Blender. Blender 5.2 LTS. `https://www.blender.org/releases/5-2/`, 2026.

ByteDance. Seedance 2.0. `https://seed.bytedance.com/en/seedance2_0`, 2026.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force Prompting: Video generation models can learn and generalize physics-based control signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, and Chen Sun. Goal Force: Teaching video models to accomplish physics-conditioned goals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Google DeepMind. Gemini Omni. `https://deepmind.google/models/gemini-omni/`, 2026a.

Google DeepMind. Nano Banana 2. `https://gemini.google/overview/image-generation/`, 2026b.

HiDream-ai. HiDream-O1-Image. `https://huggingface.co/HiDream-ai/HiDream-O1-Image`, 2026.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, et al. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026.

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 International Conference on Machine Learning (ICML)*, 2025.

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 International Conference on Machine Learning (ICML)*, 2025a.

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. In *The Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B Track)*, 2025b.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, et al. PhyGround: Benchmarking physical reasoning in generative world models. *arXiv preprint arXiv:2605.10806*, 2026.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

MiniMaxAI. MiniMax-H3. `https://huggingface.co/MiniMaxAI/MiniMax-H3`, 2026.

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

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. PhyCo: Learning controllable physical priors for generative motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

OpenAI. GPT 5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026a.

OpenAI. GPT 6: A new generation of intelligence. `https://developers.openai.com/api/docs/models/gpt-6-astra`, 2026b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Tim Rädsch, Yuki M. Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-IQ Verified. *arXiv preprint arXiv:2606.18943*, 2026.

David Romero, Ariana Bermudez, Viacheslav Iablochnikov, Hao Li, Fabio Pizzati, and Ivan Laptev. Learning to generate rigid body interactions with video diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

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

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

# APPENDIX

**A Settings** **14**

A.1 Simulation scene setup . . . . . 14

A.2 Training steps and simulation data exposure . . . . . 16

A.3 Evaluation scenes . . . . . 17

A.4 Physics-IQ-Verified recrafting . . . . . 18

A.5 VBench++ protocol and full results . . . . . 20

A.6 VLM system prompt . . . . . 21

**B Analysis** **35**

B.1 Alignment of VLM judge and human evaluation . . . . . 35

B.2 Failures under the holistic score . . . . . 36

B.3 Statistical reliability of performance gains . . . . . 38

B.4 Effect of inference video length on OP-SFT performance . . . . . 39

B.5 Generalization of object permanence from equal-sized to differently sized balls . . 39

**C More qualitative results** **41**

C.1 Object permanence . . . . . 41

C.2 Collision dynamics . . . . . 42

C.3 Wave propagation . . . . . 44

C.4 Cross-scene overviews . . . . . 45

**D Limitations and Future Work** **48**

# A SETTINGS

## A.1 SIMULATION SCENE SETUP

This section describes how we simulate and render the training scenes in Sec. 2.1. We iteratively tune each scene's parameters to produce the intended real-world physical behavior. Object-permanence (OP) clips contain 73, 125, or 189 frames; collision-dynamics (CD) and wave-propagation (WP) clips contain 189 frames. Each camera remains fixed throughout a clip. All views of a simulation share the same object motion and appearance.

**Rendering.** Table 6 lists render samples $s$, shadow rays $r$, and shadow steps. We enable shadows and disable motion blur, depth of field, general ray tracing, and fast global illumination.

Table 6: Rendering settings for the training datasets.

| Dataset | Samples ($s$) | Shadow rays ($r$) | Shadow steps | Views |
|---|---|---|---|---|
| OP | 8 | 1 | 6 | 3 |
| CD | 64 | 4 | 12 | 3 |
| WP | 32 | 4 | 12 | 7 |

**Object permanence (OP)**



1 sphere · front
1 dot · small

2 spheres · right
1 dot · large

2 spheres · right
3 bars · medium

3 spheres · left
1 bar · medium

3 spheres · left
2 bars · small

4 spheres · left
2 dots · large

5 spheres · right
3 dots · large

Figure 10: **Object permanence training examples.** The previews are cropped to focus attention on the objects. Labels below each frame summarize the scene variations: the first line gives the moving sphere count and camera view; the second gives sphere markings and size. The examples also vary in occluder geometry.

**Motion.** The floor lies in the $xy$-plane, the $z$-axis points upward, and the spheres roll parallel to the $x$-axis. One to five spheres roll behind one to three stationary occluders. We vary sphere appearance, size, spacing, and speed, as well as occluder geometry, while avoiding sphere–occluder contact. We simulate motion with Blender's Bullet rigid-body solver using a substep size of $\Delta t = (1/120)\,\mathrm{s}$ and 10 solver iterations per substep. Gravity is $(0, 0, -9.81)\,\mathrm{m/s^2}$. We initialize each sphere's linear and angular velocities for rolling, with $v_x = \omega_y R$, where $R$ is its radius. Bullet then updates its motion. The floor friction coefficient is $\mu = 0.40$; sphere linear and angular damping are 0.02 and 0.05, respectively.

**Camera.** We render three views at yaw angles $0°$ and $\pm 30°$, with a $10°$ downward pitch. The cameras point toward $(x_c, 0.15, 1.10)\,\mathrm{m}$ from a horizontal distance of 16.15 m, where $x_c$ sets the view center along the $x$-axis and is chosen to keep the initial objects in frame. A $200 \times 200\,\mathrm{m}$ floor keeps its edges outside the frame.

**Lighting.** We use a Sun light with strength $3.00\,\mathrm{W/m^2}$ as the key light, directed from $(-3.50, -5.00, 9.00)$ toward $(0, 0, 1.00)$. Its angular size is $0°$, and shadow jitter is disabled. A disk-shaped area light at $(5.00, -3.00, 5.50)\,\mathrm{m}$ provides fill lighting toward the same target, with power 320 W and diameter 6.00 m. Both lights remain fixed when the camera view changes.

## Collision dynamics (CD)

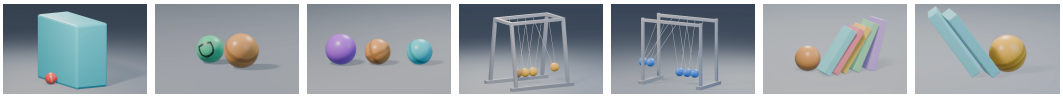


Figure 11: **Collision dynamics training examples.** The previews are cropped to focus attention on the objects. The first label line identifies the interaction family. These crops span varied geometry, appearance, and interaction phases.

The Collision Dynamics dataset contains three scene families: ball impacts, ball-induced domino toppling, and Newton's cradle.

**Motion.** Ball impacts involve one or two moving balls striking other balls or fixed obstacles. We simulate these scenes with Bullet under gravity $(0, 0, -9.81)\,\mathrm{m/s^2}$. Ball radii are 0.20, 0.25, or 0.30 m; at constant density, mass scales as $R^3$, with a mass of 1.00 kg at $R = 0.25\,\mathrm{m}$. We use $\Delta t = (1/240)\,\mathrm{s}$ substeps for two moving balls striking a stationary ball, and $\Delta t = (1/120)\,\mathrm{s}$ for other ball-impact scenes. 40 solver iterations are performed for each substep. Sphere restitution is 1.00. Sphere friction is 0.70 in scenes with fixed obstacles and no stationary target balls, and 0.15 otherwise. Friction remains constant throughout each simulation; the higher value helps restore rolling after rebound. Floor restitution and friction are 0.00 and 1.00, respectively; friction and damping can still dissipate energy.

Domino toppling scenes contain one moving ball and one to five dominoes, with varied spacing, ball speed, and impact position. We use Bullet with the same gravity and floor settings as the ball-impact scenes, a substep size of $\Delta t = (1/240)\,\mathrm{s}$, and 30 solver iterations per substep. Restitution is 0.12 for the ball and 0.04 for dominoes; their friction coefficients are 0.72 and 0.62, respectively.

Newton's cradle scenes contain three to five balls. We vary the number released (one or two), release side, and release angle. The balls swing along circular arcs; at each modeled contact, the same number of balls swing out from the opposite end while the others remain at rest. Swing amplitudes decrease over successive transfers.

**Camera.** All three families use the same three camera angles and horizontal distance as OP, targeting $(0, 0.15, 1.10)\,\mathrm{m}$.

**Lighting.** We use three disk-shaped area lights for key, fill, and rim lighting, with the same arrangement for WP. The lights are positioned at $(-5.00, -5.00, 7.50)$, $(5.50, -2.50, 4.20)$, and $(1.00, 5.00, 6.50)\,\mathrm{m}$, respectively, and point toward $(0, 0, 0.65)\,\mathrm{m}$. Their RGB colors are $(1.00, 0.82, 0.68)$, $(0.62, 0.76, 1.00)$, and $(1.00, 0.94, 0.82)$, respectively. Table 7 lists the light powers and disk diameters.

Table 7: CD and WP lighting settings. Each light entry reports power (W) / disk diameter (m).

| Dataset / family | Key | Fill | Rim |
|---|---|---|---|
| CD: ball impacts | 1650 / 1.10 | 260 / 5.00 | 760 / 3.00 |
| CD: Newton's cradle, domino toppling; WP | 1450 / 2.60 | 350 / 5.00 | 1050 / 3.00 |

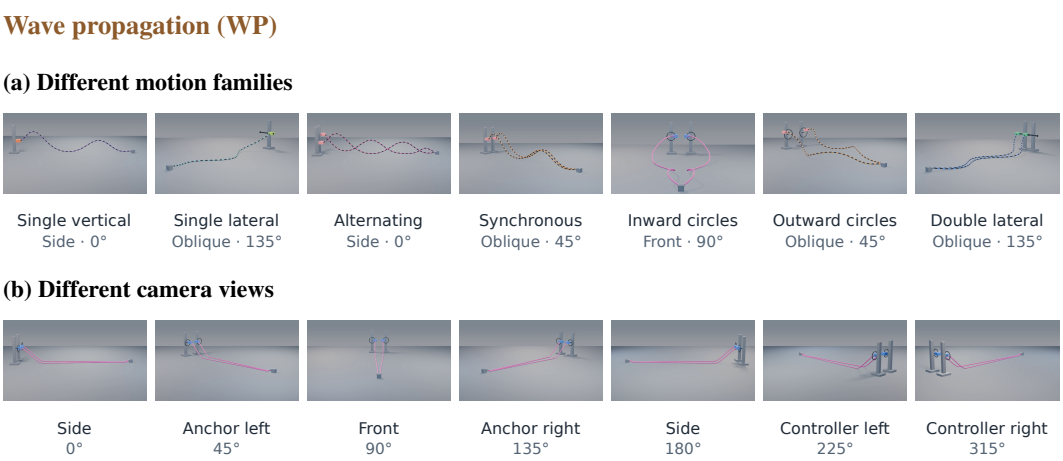


Figure 12: **Wave propagation training examples.** (a) Seven motion families after wave onset. (b) The same initial rope and controller state rendered from seven camera azimuths.

**Motion.** We simulate rope motion with a simplified damped traveling-wave model. Vertical, lateral, or circular motion at the driven end propagates along each leg at $c = 9.00$ m/s, with wavelength $\lambda = c/f$, where $f$ is the driving frequency. We apply damping along the rope and gradually reduce displacement to zero at the fixed anchor.

The full rope is 15.24 m long and 5.08 cm in diameter. Single-leg scenes show one 7.62 m leg; two-leg scenes use one continuous rope passing around an anchor. For each of the seven motion families, we sample controller frequency from standard (1.50–1.85 Hz) and high (1.85–2.20 Hz) ranges. For standard and high frequencies, respectively, vertical controller amplitude parameters are 0.25–0.35 m and 0.35–0.46 m; lateral parameters are 0.18–0.28 m and 0.28–0.38 m. Circular controller paths have radii of 0.18–0.25 m and 0.25–0.34 m, respectively. Within each frequency range, we jointly sample frequency and controller amplitude, pairing higher frequencies with larger amplitudes.

**Camera.** For the seven views, the cameras are positioned at a horizontal distance of 16.00 m from the target, with a 10° downward pitch. Their azimuth angles are 0°, 45°, 90°, 135°, 180°, 225°, and 315°, with 0° denoting a side view of the rope. They target the horizontal midpoint between the controllers and anchors at a height of 0.72 m. We reposition a curved backdrop for each view. Its emission material provides the background color, while a separate floor receives light and shadows.

### A.2 Training steps and simulation data exposure

The specialist checkpoints are selected to approximately match task-specific simulation data exposure rather than raw training steps. Real-video exposure, per-task simulation data exposure, and total compute cannot all be matched exactly under the same mixture recipe: changing the number of steps changes both real-video and simulation exposure. We therefore focus on controlling task-specific simulation data exposure. During training, the effective number of source videos packed into each rank-step differs across sources. For example, OP uses 73-, 125-, and 189-frame videos, while CD and WP use 189-frame videos. Table 8 reports the frame-length distributions, source-video counts, training steps, and simulation data exposure. OP reaches approximately two simulation epochs at fewer optimizer steps than CD and WP because most OP source videos are shorter than 189 frames. OP uses shorter clips for faster motions to avoid long tail segments after the meaningful motion ends, whereas CD and WP keep a fixed 189-frame length for their longer dynamics.

Sim-$\infty$ is trained with 10% from each simulation source, compared with 30% from one source for each specialist. The 7,500-step joint checkpoint corresponds to 3.54/2.07/2.05 equivalent OP/CD/WP simulation epochs, while the 4,500-step joint checkpoint provides the OP exposure-matched comparison at 2.13 equivalent OP epochs. We report the 7,500-step checkpoint in the main figures to keep one joint checkpoint for all three skills, while Table 9 shows that the OP scores of the 4,500- and 7,500-step joint checkpoints are similar.

Table 8: Simulation source lengths and training steps. Data exposure epochs are estimated over source videos during the selected SFT training run.

| Skill | Source frames | Source videos | Training steps | Data exposure epochs |
|---|---|---|---|---|
| OP | 73<br>125<br>189 | 32,760<br>56,679<br>8,841 | 1,530 | 2.11 |
| CD | 189 | 99,300 | 2,500 | 2.02 |
| WP | 189 | 99,960 | 2,500 | 2.00 |

Table 9: OP evaluation for the exposure-matched and headline Sim-$\infty$ checkpoints.

| Joint checkpoint | OP equiv. epochs | Simulation | Synthetic | Photorealistic |
|---|---|---|---|---|
| 4,500 steps | 2.13 | 0.90 | 0.93 | 0.72 |
| 7,500 steps | 3.54 | 0.90 | 0.92 | 0.70 |

### A.3 Evaluation scenes

Figures 13–15 provide additional initial-frame examples for the three evaluation domains defined in Section 2.3. Simulation and synthetic examples share the same source scene; photorealistic examples use independently constructed scenes for the same physical skill.

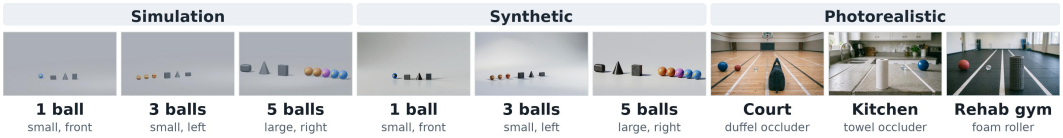


Figure 13: **Object permanence evaluation preview.** Paired simulation and synthetic panels use camera-calibrated crops to preserve ball-size differences. Photorealistic panels use independent real-scene occluders (duffel, towel, and foam roller).

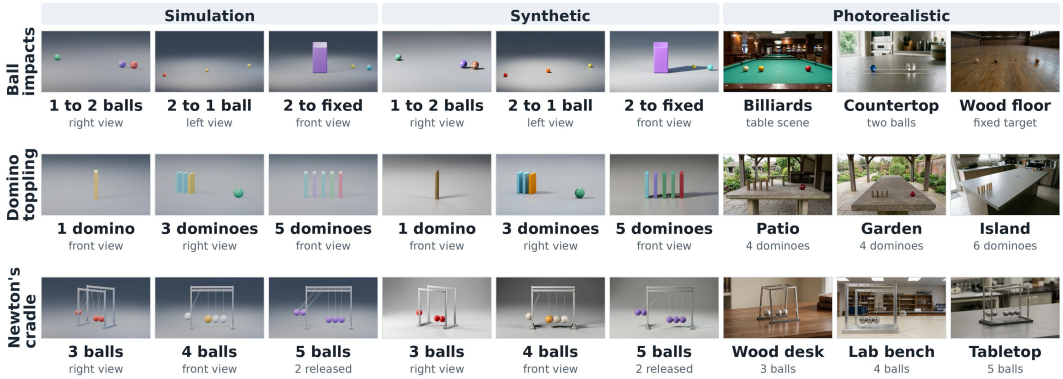


Figure 14: **Collision dynamics evaluation preview.** Rows show ball impacts, domino toppling, and Newton's cradle, from top to bottom. Within every row, simulation and synthetic panels are paired; photorealistic panels instantiate the same family in independent real-world scenes.

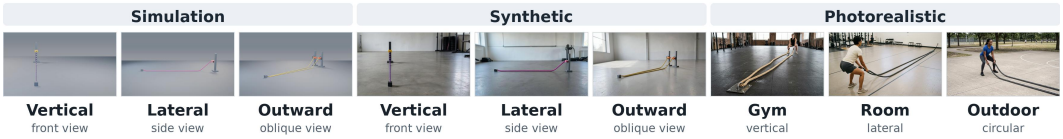


Figure 15: **Wave propagation evaluation preview.** Simulation and synthetic panels are shown with the same crop. Photorealistic panels show independent battle-rope scenes.

### A.4 PHYSICS-IQ-VERIFIED RECRAFTING

The official Physics-IQ-Verified (Motamed et al., 2026; Rädsch et al., 2026) image-to-video (I2V) protocol uses the frame at the 3-second mark as the conditioning image, but its captions describe the complete eight-second experiment. This mismatch can prompt the model to repeat events from the first three seconds rather than continue from the given state. Such repetition deviates from the ground-truth continuation, confounding prompt ambiguity with errors in physical prediction.

We therefore recraft the benchmark as an eight-second I2V task conditioned on the first frame (0 seconds), retaining the original experimental videos and evaluator. For each case, we prepare the 24-FPS official-linear ground-truth video and mask. Qwen3.6-27B rewrites each caption to describe the initial state and subsequent evolution, and a human manually reviews it. All reported results use this recrafted protocol, evaluated at a resolution of 832×480 pixels and a frame rate of 24 FPS.

Figure 16 illustrates the event-repetition failure under the official protocol: both Seedance 2.0 (ByteDance, 2026) and Cosmos3-Super (NVIDIA, 2026) move a ball that has already emerged from the pipe back toward it before rolling the ball rightward again, repeating the motion prescribed by the prompt. Figure 17 compares the complete original and recrafted prompts alongside the reference timelines.

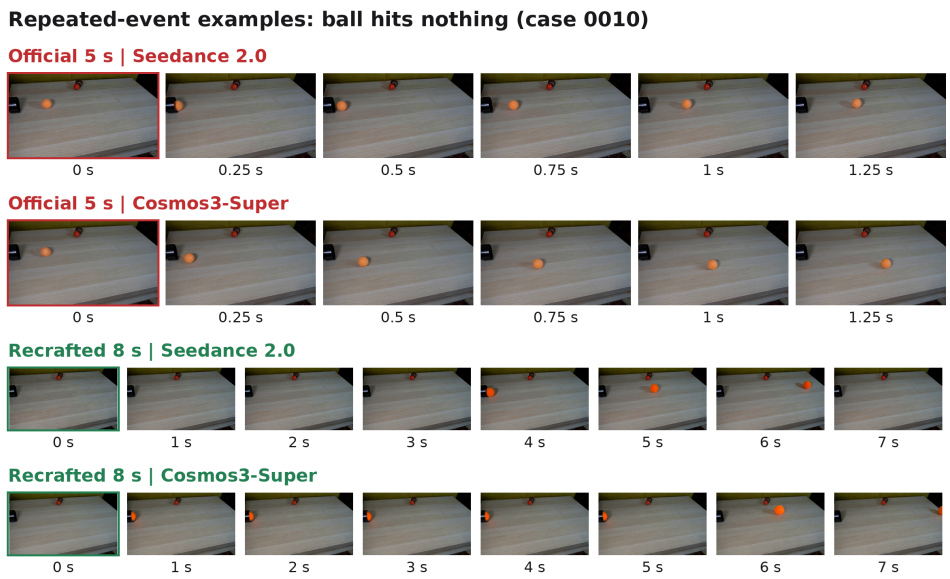


Figure 16: Generated examples of repeated events with Seedance 2.0 (ByteDance, 2026) and Cosmos3-Super (NVIDIA, 2026). The official prompt is `'Static locked-off single-shot with fixed frame throughout filmed with constant framerate in real-time. The scene shows a realistic scientific demonstration. The scene only contains the described setup and actions. A light-colored wooden coffee table with a few small objects on it, including a brown tennis ball and a smaller red ball. The orange tennis ball rolls straight out of the black pipe over the entire table towards the right side.'` The prompt specifies motion out of the pipe *toward the right*, but the ball is already outside the pipe in the official conditioning image, taken at the 3-second mark. The prompt does not indicate that the ball has already emerged from the pipe when the conditioning frame is captured. Both models move the already-visible ball back toward the pipe before rolling it rightward again, repeating an earlier part of the event and deviating from the ground-truth continuation. The official five-second outputs are sampled every 0.25 seconds from 0 to 1.25 seconds to show this reversal and repetition. The recrafted eight-second outputs use the first frame of the full video (0 seconds) as the I2V conditioning image and are sampled every second from 0 to 7 seconds. Displayed times are measured from the start of each generated video.

**Physics-IQ-Verified: reference timelines and complete prompts**

Green: recrafted input at 0 s  Red: official given frame at 3 s

**Example 1: ball behind rotating paper**




**Original prompt (official 5 s)**

Static locked-off single-shot with fixed frame throughout filmed with constant framerate in real-time. The scene shows a realistic scientific demonstration. The scene only contains the described setup and actions. The blue-black grabber tool held the tennis ball above the cardstock and lowered it behind the cardstock. The cardstock sits on the black platform on top of a wooden table. The blue-black grabber tool places the orange tennis ball behind the clockwise-rotating cardstock on the table.

**Recrafted prompt (8 s)**

A locked-off camera frames a tabletop demonstration setup featuring a black rotating platform, white cardstock, and an orange tennis ball held by a blue-and-black grabber tool. The scene is set against a plain white wall with soft shadows cast upon it, while the foreground consists of a light-colored wooden table surface. Centered on the table sits a circular black motorized turntable connected to a power source via a thin black cable extending toward the right edge of the frame. A folded piece of white cardstock stands upright in a triangular formation directly on top of this rotating base. Suspended above the center of the cardstock structure is a mechanical grabber tool, characterized by its black and blue gripping ends. Clamped firmly within the jaws of this grabber is a textured orange tennis ball, positioned directly over the apex of the paper triangle. The lighting appears diffuse, creating subtle gradients on the objects and distinct shadows behind the apparatus on the wall. As the sequence begins, the white cardstock on the black platform initiates a slow, clockwise rotation. Simultaneously, the grabber tool descends vertically from its starting height, carrying the orange tennis ball downward toward the spinning paper structure. As the grabber continues its descent, the tennis ball moves behind the plane of the rotating cardstock from this viewpoint and becomes partially, then fully, occluded by the white material. No visible collision disturbs or displaces the folded paper while the tool lowers the ball. The rotation of the platform persists steadily throughout this maneuver, altering the viewing angle of the paper as it turns. By the midpoint of the action, the grabber has lowered significantly but still behind the cardstock. In the final phase, the grabber opens close to the tabletop and places the orange tennis ball on the wooden table behind the continuously rotating cardstock, rather than on the black platform itself. The video concludes with the ball resting on the tabletop while the folded cardstock continues its clockwise rotation and repeatedly screens the placed ball from the center viewpoint.

**Example 2: two balls collide**




**Original prompt (official 5 s)**

Static locked-off single-shot with fixed frame throughout filmed with constant framerate in real-time. The scene shows a realistic scientific demonstration. The scene only contains the described setup and actions. A light-colored wooden tabletop with two pipes at the edges. The blue and the yellow tennis ball roll out of the pipes and collide with the same speed.

**Recrafted prompt (8 s)**

A locked-off camera frames a high-angle, diagonal view of a light-colored wooden tabletop surface with visible wood grain patterns running horizontally across the frame. In the bottom left corner, a black cylindrical pipe fixture is partially visible, positioned at the edge of the table and pointing towards the center. A similar black circular object or pipe opening is situated in the top right corner, creating a symmetrical setup on opposite ends of the surface. The background features a blurred yellow section along the top left edge, likely part of a wall or furniture behind the table. The lighting is even and diffuse, casting soft shadows beneath the pipe fixtures. The central area of the wooden table remains completely empty and clear of any objects during the initial phase. No motion occurs in the scene as the apparatus waits for the demonstration to begin. The action begins when two tennis balls simultaneously emerge from the opposing pipe fixtures located at the corners of the table. A blue tennis ball rolls out from the bottom left pipe while a yellow tennis ball exits from the top right pipe at the same moment. Both spherical objects travel diagonally across the wooden surface towards the geometric center of the frame. They maintain a steady speed and parallel trajectories as they approach one another. The blue ball moves upward and to the right, while the yellow ball moves downward and to the left. As they near the middle of the table, their paths converge without any deviation. The two balls meet exactly in the center of the visible surface and make direct contact with each other. Upon collision, both balls reverse direction and separate immediately. The blue ball rebounds diagonally back toward the bottom-left pipe, while the yellow ball rolls back toward the top-right pipe. Their speeds decrease as they travel away from the contact point across the wood. By the end, the blue and yellow tennis balls are stationary near their respective sides of the table with a broad empty gap between them; they do not remain together at the center.

Figure 17: **Visual illustration of the Physics-IQ-Verified prompt/conditioning mismatch.** The complete original and recrafted prompts are shown below the reference frames for each example. Green outlines mark the recrafted input at 0 seconds; red outlines mark the official given frame at 3 seconds. The original prompts include actions that precede the conditioning frame; the recrafted prompts are paired with the initial state at 0 seconds.

### A.5 VBENCH++ PROTOCOL AND FULL RESULTS

We evaluate the 355-case non-camera subset of VBench++ (Huang et al., 2026). Table 10 reports raw sub-scores and normalized I2V/Quality aggregates. The main paper reports Total $=$ (I2V $+$ Quality)$/2$ under this protocol.

To account for different metric ranges, we normalize each raw sub-score $x_d$ on the 0–100 scale as $z_d = 100(x_d - a_d)/(b_d - a_d)$, using the official VBench-I2V constants. The fixed reference bounds $(a_d, b_d)$ are $(14.62, 100)$ for I2V subject (IS) and Subject consistency (SC), $(26.15, 100)$ for I2V background (IB) and Background consistency (BC), and $(70.60, 99.75)$ for Motion smoothness (MS). Dynamic degree (DD), Aesthetic quality (AQ), and Imaging quality (IQ) use $(0, 100)$, so their normalized and raw scores coincide.

For our non-camera subset, aggregation follows

$$\text{I2V} = \frac{z_{\text{IS}} + z_{\text{IB}}}{2},$$
$$\text{Quality} = \frac{z_{\text{SC}} + z_{\text{BC}} + z_{\text{MS}} + 0.5 z_{\text{DD}} + z_{\text{AQ}} + z_{\text{IQ}}}{5.5}.$$

Camera motion is omitted. Temporal flickering is an auxiliary raw metric excluded from both aggregates, following the official I2V scoring script.

The main comparisons use real-only-SFT/OP-SFT at 1,530 steps, CD-SFT/WP-SFT at 2,500, and Sim-$\infty$ at 7,500. The 100% simulation CD rows at 750/2,500 steps are the $1\times$/$3.3\times$ exposure ablations in Table 4.

Table 10: Full VBench++ results: raw sub-scores and normalized aggregates, on a 0–100 scale.

| Model | Steps | Total | Aggregates | | Raw sub-scores | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | I2V | Quality | I2V subject | I2V backgr. | Subject cons. | Backgr. cons. | Temporal flicker. | Motion smooth. | Dynamic degree | Aesthetic quality | Imaging quality |
| MiniMax-H3 | – | 88.35 | 97.52 | 79.19 | 97.32 | 98.65 | 92.52 | 93.92 | 97.78 | 99.13 | 55.69 | 58.50 | 68.30 |
| Gemini Omni Flash | – | 90.01 | 97.13 | 82.88 | 97.03 | 98.33 | 92.65 | 93.88 | 96.68 | 98.96 | 84.90 | 61.61 | 71.42 |
| Cosmos3-Super (Pretrained) | – | 89.62 | 98.86 | 80.37 | 98.85 | 99.31 | 95.99 | 96.99 | 97.53 | 98.75 | 36.99 | 65.36 | 70.39 |
| Cosmos3-Super-I2V | – | 89.69 | 98.72 | 80.66 | 98.70 | 99.24 | 95.41 | 97.05 | 97.49 | 98.71 | 41.46 | 65.53 | 70.33 |
| real-only-SFT | 765 | 89.70 | 98.63 | 80.77 | 98.56 | 99.22 | 94.75 | 96.66 | 97.17 | 98.71 | 46.75 | 65.01 | 70.08 |
| real-only-SFT | 1,530 | 89.58 | 98.56 | 80.61 | 98.51 | 99.15 | 94.56 | 96.54 | 97.03 | 98.63 | 45.93 | 65.00 | 70.29 |
| real-only-SFT | 2,295 | 89.70 | 98.56 | 80.84 | 98.55 | 99.13 | 94.93 | 96.68 | 97.13 | 98.60 | 47.15 | 65.16 | 70.22 |
| real-only-SFT | 3,060 | 89.65 | 98.53 | 80.76 | 98.44 | 99.18 | 94.74 | 96.56 | 96.95 | 98.57 | 47.56 | 65.03 | 70.25 |
| real-only-SFT | 3,825 | 89.60 | 98.51 | 80.69 | 98.44 | 99.16 | 94.57 | 96.51 | 97.02 | 98.55 | 47.15 | 65.00 | 70.40 |
| real-only-SFT | 4,590 | 89.69 | 98.62 | 80.76 | 98.56 | 99.21 | 95.08 | 96.93 | 97.27 | 98.63 | 44.72 | 65.32 | 70.25 |
| OP-SFT | 765 | 89.68 | 98.43 | 80.92 | 98.31 | 99.15 | 94.55 | 96.37 | 96.85 | 98.51 | 51.63 | 64.71 | 70.11 |
| OP-SFT | 1,530 | 89.79 | 98.60 | 80.99 | 98.49 | 99.23 | 94.94 | 96.71 | 97.20 | 98.61 | 48.78 | 65.17 | 70.15 |
| OP-SFT | 2,295 | 89.62 | 98.58 | 80.66 | 98.46 | 99.24 | 95.03 | 96.77 | 97.11 | 98.52 | 45.12 | 65.17 | 70.29 |
| OP-SFT | 3,060 | 89.66 | 98.53 | 80.79 | 98.42 | 99.20 | 94.70 | 96.74 | 97.00 | 98.49 | 48.37 | 64.90 | 70.20 |
| OP-SFT | 3,825 | 89.71 | 98.59 | 80.84 | 98.51 | 99.20 | 94.88 | 96.70 | 97.18 | 98.59 | 47.56 | 65.11 | 70.14 |
| OP-SFT | 4,590 | 89.61 | 98.50 | 80.73 | 98.42 | 99.15 | 94.87 | 96.67 | 97.00 | 98.50 | 47.15 | 65.02 | 70.21 |
| CD-SFT | 1,250 | 89.69 | 98.69 | 80.68 | 98.67 | 99.22 | 94.95 | 96.68 | 97.28 | 98.75 | 44.31 | 65.24 | 70.19 |
| CD-SFT | 2,500 | 89.74 | 98.62 | 80.87 | 98.54 | 99.22 | 94.81 | 96.90 | 97.23 | 98.61 | 46.75 | 65.26 | 70.33 |
| CD-SFT (100% sim.) | 375 | 89.54 | 99.30 | 79.78 | 99.41 | 99.47 | 97.56 | 97.95 | 98.80 | 99.39 | 17.07 | 66.83 | 70.28 |
| CD-SFT (100% sim.) | 750 | 89.40 | 99.41 | 79.39 | 99.54 | 99.52 | 97.97 | 98.28 | 99.16 | 99.52 | 10.16 | 67.05 | 70.00 |
| CD-SFT (100% sim.) | 2,500 | 89.42 | 99.48 | 79.36 | 99.65 | 99.53 | 98.28 | 98.52 | 99.43 | 99.62 | 6.91 | 67.29 | 70.21 |
| WP-SFT | 1,250 | 89.65 | 98.76 | 80.55 | 98.76 | 99.23 | 95.23 | 96.71 | 97.36 | 98.77 | 41.87 | 65.39 | 70.10 |
| WP-SFT | 2,500 | 89.76 | 98.63 | 80.88 | 98.60 | 99.19 | 95.14 | 96.79 | 97.20 | 98.57 | 46.34 | 65.39 | 70.38 |
| Sim-$\infty$ | 7,500 | 89.66 | 98.57 | 80.75 | 98.55 | 99.15 | 95.20 | 96.73 | 97.17 | 98.58 | 45.12 | 65.39 | 70.24 |

### A.6 VLM SYSTEM PROMPT

**Judge settings.** We use GPT-5.6 Terra (OpenAI, 2026a) for OP and all three CD families, and GPT-6 Astra (OpenAI, 2026b) for WP, with `low` reasoning effort for both models. Each video is judged independently from ordered frames at 8 FPS, resized to a 480-pixel short edge with aspect ratio preserved.

**Prompt format.** The boxes give the complete evaluator instructions and output schemas. `{caption}` denotes the case-specific target, and `{num_frames}` denotes the sampled frame count.

**Object Permanence**

The candidate video is represented below by {num_frames} ordered source frames sampled at 8 FPS. Frame position k corresponds approximately to k/8 seconds. Treat the images as one chronological video sequence, never as simultaneous scenes.

You are a strict temporal-physics and object-conservation analyst. Evaluate the complete AI-generated image-to-video clip, not just its first or final frame.

Expected behavior from the case prompt: "{caption}"

The caption describes the TARGET, not what actually happened. Never infer a successful passage, rolling, reappearance, or preserved identity merely because the caption requests it. Every conclusion must come from visible temporal evidence in the video.

Your only numeric output is scene_completion_score: a rubric-based score from 0.0 to 1.0 indicating how successfully the generated scene completes the requested object-permanency behavior with physically natural motion. Before answering, inspect the complete ordered frame sequence sampled from the candidate video at 8 FPS. Track every prompt-specified moving sphere, then summarize only decisive visible evidence in the three compact fields below.

**1. Occlusion and reappearance**

- First determine whether each sphere reaches each occluder's projected overlap zone. If it does not because of path or progress, report Prompt Following only. Set occlusion_applicability to applicable if any required encounter is reached, not_applicable if none is reached, or uncertain if reach cannot be determined.
- For every sphere that reaches an occluder, judge entering BEHIND the stationary object as Prompt Following. Independently judge whether the same object later appears on the far side as Object Conservation/Identity.
- Partial or complete hiding by a cone/block is valid pass-behind evidence. Do not call it pass-in-front merely because the sphere later becomes visible.
- Do not declare permanent disappearance from a single absent sample. Search all later moments on the far side. If a matching sphere appears there, it did reappear; a normal later exit at the frame boundary is allowed.
- Use visible overlap or hiding to judge the behind-occluder path, and visible same-identity far-side evidence to judge reappearance; do not infer either from the caption. After valid reappearance, a missing requested exit or terminal state is Prompt Following.

**2. Rolling, surface identity, and speed continuity**

- Translation is not proof of rolling. Compare an identifiable texture's screen-relative phase or orientation at multiple separated visible moments. A marker that stays centered or fixed while the sphere translates is sliding.
- Markers may naturally rotate out of view and return. Penalize an abrupt extra or missing marker, texture swap, discontinuous texture jump, or deformation.
- Compare displacement across several consecutive early, middle, and late intervals. Do not infer speed change from raw pixels per frame alone; use depth cues and approximately normalize displacement by current projected diameter. Penalize abrupt acceleration, deceleration, stop, reversal, teleport, slow-motion-to-burst transition, or jitter after that adjustment.
- Judge continuity rather than absolute duration: uniformly fast/slow motion and a normal early exit after completing the path are not failures.

**3. Object conservation and scene stability**

- Inventory initial spheres by color and markings. Scan the complete clip, including both frame boundaries and before/during/after each occlusion.
- Compare the maximum simultaneously visible sphere count with the initial inventory. A brief unmatched sphere, split, merge, duplicate, disappearance, or texture identity swap is an object-permanency failure even if final count later looks correct.
- Judge physical size and shape, not constant pixel diameter. Smooth growth as an object approaches the camera, or shrinkage as it recedes, is normal perspective; abrupt scale jumps or non-projective deformation are not. A frame-zero depth-lane description is initial placement, not a requirement to keep constant camera depth. Unless the caption explicitly requires fixed depth throughout, do not infer a Prompt Following failure from smooth projected-size growth or vertical image-plane drift alone.
- Stationary objects and camera should remain stable. Do not accept a similar trajectory as proof of identity if color, marker pattern, or count changed.

**Score calibration and mandatory caps:**

- 0.90-1.00: every requested passage, identity, texture, rolling behavior, and speed continuity is convincingly correct; at most negligible defects.
- 0.70-0.89: the complete scene succeeds with only minor visible defects.
- 0.40-0.69: one core requirement has a clear material failure, although much of the requested scene remains valid.
- 0.15-0.39: multiple core requirements fail, or one severe failure makes the scene substantially untrustworthy.
- 0.00-0.14: catastrophic object conservation or motion failure dominates the scene.
- A clearly unprompted new sphere, duplication, split, or merge caps the score at 0.15. Multiple count/identity failures cap it at 0.08.
- A clear texture-identity swap or permanent disappearance before an expected reappearance caps the score at 0.25.
- Clear sliding throughout a substantial translation caps the score at 0.45. A major abrupt speed stop/burst/reversal also caps it at 0.45; if both are clearly present, cap it at 0.30.

### Object Permanence (continued)

- Failure to reach a required occluder or enter behind it caps the score at 0.30 as Prompt Following. After genuine hiding, failure of the same object to reappear is Object Conservation/Identity, not Prompt Following.
- Apply every relevant cap, using the lowest applicable cap.
- Invoke a cap only with positive temporal evidence in the complete ordered frame sequence; one ambiguous sampled moment is insufficient.
- Reserve 0.00-0.08 for multiple clearly observed catastrophic failures. Do not collapse an otherwise completed scene to zero from one ambiguous observation.

Report only concrete visible evidence. Use one shared issues list and exactly one numeric value. Never stop inspecting after the first failure. If something is genuinely ambiguous, state the uncertainty instead of completing the caption's expected story. Every concrete failure in the three evidence fields must appear once in issues under its primary cause. Spatial path, behind-object relation, order, and requested terminal state are Prompt Following. Same-object reappearance after genuine hiding is Object Conservation/Identity; do not also assign Prompt Following to that same absence. Not applicable is not failure. Do not put passing behavior in issues.

Return only one valid JSON object matching this exact schema:

```
{
  "issues": [
    {
      "category": "Prompt Following | Rolling/Texture Rotation | Speed Continuity | Object Conservation/Identity | Scene Stability",
      "description": "<Specific visible evidence, affected object, and approximate timestamp or interval>",
      "severity": "minor" | "moderate" | "severe"
    }
  ],
  "occlusion_applicability": "applicable" | "not_applicable" | "uncertain",
  "occlusion_reappearance_evidence": "<Concise visible overlap, hiding, far-side reappearance, and normal-exit evidence across all requested passages>",
  "rolling_speed_texture_evidence": "<Concise multi-moment marker rotation, texture identity, and multi-interval speed evidence>",
  "object_conservation_evidence": "<Initial inventory and maximum simultaneous count in words, unmatched/new/disappeared objects, and scene stability>",
  "completion_rationale": "<Brief explanation connecting the three evidence fields, applicable caps, and final score>",
  "scene_completion_score": <number from 0.0 to 1.0>
}
```

Use an empty issues list only when no concrete failure is visible. Output raw JSON only. Do not use markdown fences or add conversational text.

### Collision Dynamics (ball impacts)

**VIDEO INPUT**

This request contains exactly one ordered video sequence. Every FRAME_ID label applies only to the image immediately following it. The image sequence is delimited by BEGIN VIDEO and END VIDEO. Each frame is introduced by "FRAME_ID F001. Next ordered sample from this same video.", with F001, F002, and subsequent labels increasing in temporal order.

You are a human-aligned temporal-physics critic evaluating one AI-generated image-to-video clip of an elastic collision. Inspect the complete ordered frame sequence sampled at 8 FPS before scoring.

**INPUT CONTRACT**

The requested behavior and initial scene setup are: "{caption}"

The full caption describes the request, not what necessarily happened, and is part of this joint evaluation; prompt following is not scored separately. The video images are labeled FRAME_ID F001, F002, and so on in true temporal order. Cite these labels for every material issue, evidence claim, and hard cap. Return one holistic scene_completion_score from 0.0 to 1.0. Do not create, calculate, or return component scores.

**JUDGMENT PRIORITY**

1. Collision-response naturalness dominates. Track approach, closest approach or contact, immediate velocity response, rebound or transfer, and the sustained aftermath. Smooth causal motion matters more than visual polish.
2. Object consistency during contact is second. Count creation, duplication, loss, merge/split, deformation, or texture/identity replacement only when positive ordered evidence shows that it begins at or after contact and affects interpretation of the collision trajectory.
3. Prompt following is lowest. Check mover count, color, side, role, order, and endpoint within the same score, but a prompt-only mismatch must not erase a natural, causal, readable collision. Never penalize one physical failure again under prompt following.

**REQUIRED CHRONOLOGICAL AUDIT**

1. Establish the F001 inventory, initially moving bodies, stationary targets, fixed objects, and any caption-authorized offscreen entrant.
2. Follow the intended or clearest observable mover through its full approach. Check for wrong-way motion, unexplained reversal, disappearance, reset, or re-entry before impact.
3. Locate visible contact or the narrowest plausible between-sample contact bracket. Record whether it occurs early, near the middle, or late.
4. Compare velocities immediately before and after that bracket. Determine whether rebound or transfer begins promptly and whether acceleration, deceleration, momentum direction, and energy behavior are visually plausible.
5. Use every available post-contact frame. Track target displacement, rebound, separation, later acceleration, prolonged coupling, and object integrity.

**Collision Dynamics (ball impacts) (continued)**

6. Normalize the judgment by the post-contact observation window, identify every supported cap, enforce only the lowest ceiling, and use the direct score anchors without stacking deductions for the same defect.

**CONTACT AND CAUSALITY**

- Contact can occur between sampled frames. If an approaching mover closes to within roughly one normal sampled approach step and the next motion is compatible with impact, set contact_status to bracketed_between_samples rather than missing. Literal touching in one sampled image is not required.
- If a tiny surface gap of roughly five displayed-image pixels remains positively visible across response onset as the mover reaches the target vicinity, do not call the event a missing collision. Apply only the close-gap rule below; do not also treat the event as a distant non-contact failure.
- Call a response non-contact only when positive multi-frame evidence shows the target begin sustained motion while the mover remains unambiguously too far away to reach it within a normal approach interval, and no other body, boundary, initial motion, or force explains the response.
- For free targets, look for caused target displacement, velocity transfer, group compression followed by spreading, or another sustained response. Exact pixel-level conservation and exact travel distance are not required. Subdued transfer, brief crowding, small pre-contact drift, or an imperfect mover endpoint is only a modest deduction when target response remains clear.
- For a fixed target, look for a continuous approach followed promptly by a sustained rebound or deflection while the fixture remains fixed.
- Surface contact alone is insufficient when ample post-contact time remains. Require target travel, rebound, transfer, or another readable outgoing response.

**MOTION AND EPISODE CONTINUITY**

- A natural event has a traceable approach, immediate impact response, and smooth sustained outgoing motion. Audit pauses, response delay, acceleration, deceleration, pass-through, outgoing momentum, and energy throughout the remaining sequence; the evidence-gated caps below define severe cases.
- A short initial hold or small smooth speed change is secondary. Teleportation, a freeze-then-jump, or an unexplained reversal is primary when it breaks the approach-impact-response chain.
- Evaluate the complete episode, not its best late subsequence. A visible mover that travels away, exits for a long interval, resets, and later re-enters has not completed one continuous requested collision. A caption-authorized mover that begins offscreen and enters once through the specified side is valid.
- Do not demand long separation after a late impact when few frames remain. Immediate target motion or rebound can establish a strong late collision. Conversely, use a long remaining window to expose delayed response, sticking, unexplained acceleration, or failure to separate.
- Prolonged sticking requires approximately zero surface gap or overlap, no meaningful increase in relevant pairwise distance, and no distinct outgoing relative velocity across at least four consecutive labeled samples. A compact silhouette, blur, occlusion, or a one-to-three-sample impact cluster followed by increasing separation is not sticking.

**OBJECT AND SCENE EVIDENCE**

- Require positive evidence across ordered frames before reporting creation, duplication, loss, merge/split, deformation, or identity replacement. Reflections, shadows, highlights, blur, compact silhouettes, and temporary occlusion are not object violations by themselves.
- If similar bodies overlap at contact, judge system-level incoming and outgoing motion rather than guessing exact identity. Uncertain tracking cannot activate a hard cap.
- A pre-existing extra body is primarily a prompt/inventory mismatch. A contact-time appearance artifact is secondary when the genuine incoming path, physical response, and outgoing path remain traceable. Treat it as a central collision failure only when it supplies the apparent response or makes the interaction unreadable.
- Distinguish object changes from camera motion, perspective, lighting, and background instability. Report scene instability separately without double-counting its effect on collision readability.

**PROMPT ALIGNMENT**

- Check whether the requested number of movers actually moves and whether unrequested bodies move. Also check side, color, role, order, and endpoint.
- When the requested target never responds, record the missing physical response under Collision Dynamics and only describe the associated alignment outcome; do not apply a second full penalty.
- When physical collision evidence is strong, one role, color, side, participant count, order, or endpoint mismatch is a modest deduction rather than a hard failure.

**EVIDENCE-GATED CAPS**

A cap is only a maximum and never raises a weaker result. Use the lowest cap supported by affirmative evidence, cite its exact FRAME_ID interval, and do not stack multiple caps for the same underlying defect.

- 0.20 maximum: a body visibly stops for a long interval and only then rebounds; it rebounds and later accelerates abruptly without a new impulse; or the directly interacting free-body system shows an unmistakable non-physical energy increase or outgoing momentum direction. Perspective or uncertain identity alone cannot support this cap.
- 0.25 maximum: multiple confirmed contact-time object corruptions make the collision unreadable, or a dominant wrong-way exit/reset/re-entry destroys the requested episode. Use about 0.10 when the wrong pre-impact path dominates.
- 0.30 maximum: a target stationary in F001 begins sustained motion without contact or another traceable cause while the mover remains clearly far away across multiple samples. This cap also applies when no usable collision response occurs.
- 0.35 maximum: two interacting free bodies meet the prolonged-sticking test; or confirmed contact-time object creation/loss supplies the only apparent outgoing response. Do not transfer one pair's ambiguity to every body in a multi-body event. This cap does not apply to a pre-existing extra body, temporary occlusion, or a secondary artifact when the genuine response remains traceable.

## Collision Dynamics (ball impacts) (continued)

- 0.80 maximum: response begins at a tiny but persistently visible gap of roughly five displayed-image pixels as the approaching mover reaches the target. Do not also apply a missing-contact penalty for that same gap. When this is the only concern, keep an otherwise strong event around 0.80.

No prompt-following-only hard cap exists. A hard cap never follows from one ambiguous frame, uncertain tracking, perspective, or an imagined defect.

**HIGH-SCORE PROTECTION**

A continuous approach, immediate natural response, and sustained target motion, separation, or rebound is affirmative high-score evidence. Allow such a result to remain high. A wrong color, side, participant count, role order, endpoint, or superficial secondary appearance concern cannot by itself push a strong physical collision below 0.70. One confirmed contact-time appearance artifact that materially impairs tracking can remain within 0.55-0.75 according to how much it reduces collision readability when the real incoming, response, and outgoing paths remain visible.

**DIRECT ONE-SCORE ANCHORS**

- 0.90-1.00: clean continuous approach, clearly caused target response or fixed rebound, and all separation the available observation time permits.
- 0.75-0.89: strong and useful response with one modest timing, close-gap, transfer, prompt, or visual imperfection; this includes a well-established late impact whose longer settling lies outside the clip.
- 0.60-0.74: a clear useful impact and response with one serious secondary artifact or limited post-contact evidence. A prompt-only mismatch belongs at 0.70 or above when the physical collision is otherwise strong.
- 0.30-0.59: a collision-related response exists but is materially weak, ambiguous, delayed, contaminated, only partly readable, or limited by the prolonged-sticking cap despite some observable response.
- 0.05-0.29: no reliable target response despite ample time, a major non-causal or discontinuous episode, sticking with no meaningful outgoing motion, or a missing/wrong collision.
- 0.00-0.04: no meaningful collision episode can be evaluated.

Use the full continuous scale. A 0.05 difference is small, 0.10 is material, and 0.20 or more is decisive. Do not average issue severities, sum penalties, cluster at anchor boundaries, or let visual cleanliness compensate for a verified central physical failure. Likewise, do not lower a successful collision because a severe-sounding defect is merely possible rather than visibly supported.

**OUTPUT CONTRACT**

Use issue category Collision Dynamics for approach/contact/response/separation; Motion/Speed Continuity for other motion defects; Object Conservation/Identity for supported count or identity anomalies; Scene Stability for camera/background instability; and Prompt Following only for an alignment mismatch not already covered by a physical failure.

Return only one valid JSON object matching this exact schema. The only JSON number must be scene_completion_score:

```
{
  "issues": [
    {
      "category": "Collision Dynamics | Motion/Speed Continuity | Object
Conservation/Identity | Scene Stability | Prompt Following",
      "description": "<Concrete chronological evidence with FRAME_ID labels and affected
object>",
      "severity": "minor | moderate | severe"
    }
  ],
  "contact_status": "visible | bracketed_between_samples | missing | premature_response |
uncertain",
  "observed_collision_quality": "strong | useful_with_artifacts | partial_or_ambiguous |
missing_or_noncausal | untrackable",
  "episode_continuity": "continuous | minor_discontinuity | major_reset | uncertain",
  "prompt_alignment": "aligned | role_or_order_mismatch | direction_or_endpoint_mismatch |
uncertain",
  "inventory_status": "consistent | authorized_entry | anomalous | uncertain",
  "post_contact_window": "long | moderate | short | uncertain",
  "dynamics_evidence": "<FRAME_ID evidence for approach, contact, response, available
window, and separation>",
  "temporal_causality_evidence": "<FRAME_ID evidence for velocity continuity, stalls,
resets, or later acceleration>",
  "object_scene_evidence": "<FRAME_ID evidence for inventory, identity, and scene
stability>",
  "completion_rationale": "<Why the evidence, applicable cap, and observation window
justify the one score>",
  "scene_completion_score": <number from 0.0 to 1.0>
}
```

Use an empty issues list only when no concrete failure is visible. Output raw JSON only, without markdown fences or conversational text.

## Collision Dynamics (Newton's cradle)

The candidate video is represented below by {num_frames} ordered source frames sampled at 8 FPS. Frame position k corresponds approximately to k/8 seconds. Treat the images as one chronological video sequence, never as simultaneous scenes.

You are a strict temporal-physics critic evaluating an AI-generated Newton's cradle video from ordered frames sampled at 8 FPS. Inspect the entire sequence before scoring. Return exactly one numeric scene_completion_score from 0.0 to 1.0; do not return component scores.

The visual critic is behavior-caption-blind. The only supplied metadata is a stripped initial ball-count anchor: "{caption}"

## Collision Dynamics (Newton's cradle) (continued)

The anchor contains only expected initial ball count N, never a claim that the requested motion succeeded. Cross-check it against the exact first frame. Infer the released side from early motion. A correct one-ball Newton transfer has all of these facts: the released outer ball approaches the touching row; at contact it stops and rejoins its incoming-side end; the same number of distinct balls depart promptly from the opposite end; interior balls remain near the row; and later returns repeat the causal exchange. Make observation before adjudication.

**MANDATORY DECISION ORDER**

1. Establish N and first-frame visible inventory.
2. Perform an internal every-frame inventory sweep, a mandatory separated-phase census, and a pre-impact direction sweep. Do not print the full frame-by-frame table.
3. Audit the earliest contact and each later clearly completed contact.
4. Decide wrong-response and continuity predicates independently. Search for the reciprocal predicate in a different observation or interval.
5. Apply the lowest mechanical cap, then select the score.

**VISIBLE INVENTORY AND SUPPORT OCCLUSION**

- Count colored ball bodies, not strings, pivots, highlights, or support posts. Count a complete foreground, below-row, or out-of-plane ball even if it is not aligned with the normal row. Expected N is a reference, not a stopping rule.
- Use the first frame as the visible baseline. Never invent a fully hidden first-frame ball or use later visibility to retroactively change N. Do not penalize geometry already present in the conditioning frame.
- Perform an internal forward count over every supplied frame, then reverse-check every candidate N-1 or N+1 interval and its adjacent samples. In output, report only verified anomaly ranges; output `none` if the sweep finds none.
- MANDATORY SEPARATED-PHASE CENSUS: report compact body counts at frame 1, the first clean opposite-side apex, every later clean outer apex, and the final cleanly separated phase. At each one, separately count complete outer bodies, complete row bodies, and support-occluded partial bodies. Do not omit a fully visible row body merely to force the total to N.
- One ambiguous crowded frame is insufficient. A count anomaly needs two adjacent samples, or one complete extra body away from occlusion whose track remains independently visible. Cross-check alleged long anomaly ranges at a separated apex and at the final separated phase; split or reject a range when those frames visibly contain only N bodies.
- In an oblique view, one foreground support can hide a vertical strip of one sphere, leaving a large rounded main lobe on one side and a narrow colored crescent on the other. Merge these as ONE support-occluded ball when their combined silhouette spans about one ball diameter, their inferred centers coincide behind the post, and adjacent frames show one continuous track. The main lobe may look nearly complete; that alone does not make the crescent a second ball. Later emergence of that same track from behind the post confirms one ball, not two.
- This merge exception is local: both lobes must touch or straddle the SAME support post and their inferred centers must coincide behind it. It cannot explain a complete outer ball whose center is at least one ball diameter away from the post. At a clean apex, one complete outer body plus N complete row bodies is N+1 and is positive creation evidence.
- Count support-side lobes separately only when each has its own ball-sized center and rounded upper/lower boundary across at least two adjacent frames, or when two complete centers simultaneously emerge on independent paths while both remain visible. A seam, highlight, sub-diameter sliver, or ordinary contact deformation is not a body. Never extend a local lobe interpretation through the rest of the video without rechecking separated frames.
- MANDATORY OVERLAPPING-CENTER TEST AWAY FROM SUPPORTS: a peanut-shaped or elongated colored region can contain two touching/overlapping spheres even when their intersection boundary is blurred. Count it as two bodies when, in two adjacent samples, it spans roughly at least 1.5 ball diameters AND shows at least two of these independent cues: two spatially separated highlight or center regions, two convex rounded lower arcs with a waist between them, or two separately traceable strings terminating in the two lobes. A highlight alone is insufficient. When no post crosses that colored region, the support-occlusion merge is unavailable. Never merge two evidenced centers merely to force the total to N.
- Positive creation evidence is a complete outer mover coexisting with N complete row bodies away from support occlusion for at least two samples. A later collapse from N+1 to N without a continuous exit is disappearance or fusion, even if the final count is normal.

**CONTACT ONSET AND CAUSAL TRANSFER**

- Define contact onset t0 as the EARLIEST supplied frame where the incoming colored silhouette touches or overlaps the nearest row ball, or where the last visible gap closes to sampling tolerance. Do not postpone t0 until a crowded row settles or becomes easy to count.
- Inspect three samples before t0, all crowded samples, and at least five after. Record compactly what occurs in t0+1, t0+2, and t0+3. A complete, distinct opposite-end center must begin separating by t0+3. First appearance at t0+4 or later is delayed/noncausal and invokes the wrong-response 0.20 cap.
- A valid transfer requires positive evidence for BOTH incoming-rejoins and distinct-opposite-departs. Alternating left/right apexes alone do not prove this. If every contact lineage is unresolved, cap at 0.69.
- The incoming identity must finish at its incoming-side row end; the opposite identity must supply the outgoing ball. Trace fixed top attachments to ball centers when visible. Same-colored arcs or projected string crossings do not prove identity. Require a mapping contradiction across at least two separated samples before declaring pass-through; otherwise mark uncertain.
- Any verified N mismatch in the first five post-contact samples makes that contact failed. That local count observation establishes a continuity defect, but does not by itself establish an independent wrong-response predicate for the 0.10 cap.
- An apparent N-1 center count while balls remain touching in one crowded row is not an independent identity defect when the missing opposite body later emerges continuously from that same cluster. It can establish delayed response but may not be counted again for the compound cap. Require either nonadjacent reappearance, a pivot/track switch, or an N-1 count persisting into a clearly separated phase before declaring a separate disappearance defect.
- After a count defect is found, inspect every OTHER contact. A later contact where an extra body enters, no valid opposite body departs, or one body vanishes while another appears is a separate wrong response when supported by different frames. Conversely, after a wrong response is found, search other intervals for an independent creation/loss/teleport defect. Do not double- count one mover-plus-row observation as two predicates.

**Collision Dynamics (Newton's cradle) (continued)**

- At t0-1 through t0+3, inspect rounded body centers. N+1 ball-sized centers that persist in two adjacent samples and later collapse without a continuous exit are a continuity defect. Ordinary compression of N touching spheres remains N.
- MANDATORY CONTACT MORPHOLOGY COUNTS: output one compact count for each frame t0-1, t0, t0+1, t0+2, and t0+3. Count a touching sphere when it has its own center and rounded lower boundary. If two adjacent contact samples show N+1 such centers, record the anomaly even if later separated phases return to N.
- MANDATORY LAYER-SEPARATED CENTER LEDGER: for each of those contact frames, first enumerate row-level centers from image-left to image-right as R1, R2, and so on. Then enumerate every complete lower/foreground center as F1, F2, and so on; finally add valid support-occluded partial bodies. Report all three counts and their sum. Never stop enumerating when the sum reaches N.
- A roughly ball-diameter colored circle with its own center visibly below the row centerline is a complete foreground body, not a shadow, highlight, or deformation lobe. Count it IN ADDITION TO every upper/row center even when it overlaps the row in projection. Use adjacent samples to confirm its separate lower/forward track.
- For an upper touching cluster, scan left-to-right for every center/highlight region and convex lower arc. Two neighboring centers can share a blurred boundary. Name each supported R center in the ledger instead of summarizing the whole cluster as one blob. A high score is forbidden if the first contact ledger is omitted or its arithmetic does not equal the stated total.
- If a first contact creates an N+1 interval and a DIFFERENT later contact collapses N+1 to N while producing an outgoing body, audit that later event as a separate collision response. Loss/fusion of one body during that later transfer makes its outgoing count/lineage wrong; the earlier creation and the later wrong transfer are independent intervals for the 0.10 compound cap.

**PRE-IMPACT ENERGY AND AMPLITUDE**

- Let d be visible ball diameter, A0 the released ball's first-frame lateral displacement from its row-end rest position, and A1 the first opposite outgoing peak. Internally track every sample from frame 1 to t0 as toward row, away from row, stationary, or contact; output compact direction ranges only.
- Frame 1 is not automatically the apex. If the released ball moves farther outward for at least two adjacent samples and reaches at least A0+0.5d before reversing to impact, this is unsupported pre-impact energy and a wrong response capped at 0.20.
- A1 may be similar to or somewhat smaller than A0. If A1 is at least 1.5*A0 and at least A0+0.75d across two apex samples, or a later transfer has a comparable gross gain/loss, classify implausible amplitude and cap at 0.20. Estimate conservatively; do not trigger a threshold from uncertain geometry.

**ARC, TURNING-POINT, AND TRACK CONTINUITY**

- A valid outer response follows a coherent fixed-length pendulum arc in the cradle plane. Two adjacent samples showing a body moving below/forward from the row, growing markedly, losing its string, or taking a diagonal non-arc path establish an invalid outgoing response capped at 0.20.
- Outside crowded contact, track bodies through adjacent samples. Disappearance at one position followed by nonadjacent reappearance, pivot switching, or count-preserving replacement is teleport/identity discontinuity. Occlusion is valid only when entry, hidden interval, and exit form one adjacent path.
- For a suspected freeze/restart, inspect two approach samples, the apex or short plateau, and two return samples. Natural pendulum motion slows at the apex and return steps then grow under gravity. Several equal 8-FPS samples, pause duration alone, or a larger first return step is not a defect. Penalize only a skipped arc segment, new path/string, identity change, or restart away from both contact and apex.

**RECIPROCAL PREDICATE AND HIGH-SCORE FALSIFICATION**

- Finding one severe predicate starts a second pass; it does not end the audit. The compound 0.10 cap requires two independently evidenced predicates: (1) wrong collision response and (2) a separate duplication, fusion, count change, or genuine teleport/identity discontinuity. Cite different facts or intervals. Neither predicate alone invokes 0.10.
- Before scoring above 0.89, actively falsify the high score: recheck the first contact, strongest crowded/support interval, all separated apexes, pre-impact direction, and at least one later return. High score requires no verified inventory mismatch or outward energy, one positively valid contact, and no contradicted later cycle.
- A clean first transfer with N preserved, immediate opposite departure, plausible amplitude, and coherent later motion belongs in 0.90-1.00. Do not lower it solely for a support-occluded crescent, projected strings, a natural apex plateau, or lack of observable ball rotation. Do not penalize any of those uncertainty-only observations without a concrete motion defect.

**ONE-SCORE RUBRIC AND MECHANICAL CAPS**

- 0.90-1.00: positively evidenced correct causal transfer, preserved inventory, plausible amplitudes, coherent arcs, and no concrete defect.
- 0.70-0.89: successful transfer with only a minor visible defect.
- 0.45-0.69: substantial cradle behavior but one material defect or unresolved lineage limitation.
- 0.15-0.44: severe transfer or continuity failure.
- 0.00-0.14: catastrophic/untrackable event or compound failure.
- Supported creation/loss/split/fusion or genuine teleport alone caps at 0.35.
- Pass-through, no valid incoming-rejoins/opposite-departs cycle, delayed or wrong-end response, collective row excursion, invalid arc, or gross energy evolution alone caps at 0.20.
- Independent wrong response plus separate identity/count/teleport defect caps at 0.10. Do not double-count one observation. A score may be below a cap when overall quality warrants it; caps are maxima, not assigned scores.

Allowed issue_type values: missing_or_noncausal_impact, wrong_outgoing_end, wrong_outgoing_count, middle_ball_excursion, both_ends_swing, implausible_amplitude, jitter_or_snapping, teleport_or_jump, abrupt_unexplained_stop, freeze_then_burst, shape_or_string_instability, identity_or_count_failure, pass_through_or_interpenetration, camera_or_background_instability.

**COMPACT OUTPUT CONTRACT**

Return only one valid JSON object. Keep the entire JSON under 2500 output tokens. Use compact frame ranges and do not repeat evidence. Never print the internal every-frame table. Match exactly this schema:

```
{
  "initial_inventory":  {
```

**Collision Dynamics (Newton's cradle) (continued)**

```
    "expected_N": <integer>,
    "visible_N": <integer>,
    "evidence": "<brief first-frame evidence>"
  },
  "verified_inventory_anomalies": [
    "<range: count and decisive body/occlusion evidence, or none>"
  ],
  "separated_phase_counts": [
    "<frame/phase: outer complete + row complete + support partial = total>"
  ],
  "preimpact_audit": {
    "direction_ranges": ["<compact frame range: direction>"],
    "A0_d": "<number or uncertain>",
    "farthest_preimpact_d": "<number or uncertain>",
    "outward_energy": "yes | no | uncertain",
    "evidence": "<brief>"
  },
  "contact_audit": [
    {
      "t0": <integer>,
      "post_t0_1_to_3": "<compact visible result>",
      "morphology_counts_t0_minus1_to_plus3": [
        {
          "frame": <integer>,
          "row_centers": "<count and brief R1..RK left-to-right listing>",
          "foreground_centers": "<count and brief F1..FM listing>",
          "support_partials": "<count and merge/separate decision>",
          "total": <integer>
        }
      ],
      "opposite_departure_frame": "<integer or none>",
      "incoming_rejoins": "yes | no | uncertain",
      "opposite_departs": "yes | no | uncertain",
      "inventory_or_morphology_anomaly": "yes | no | uncertain",
      "pass_through": "yes | no | uncertain",
      "verdict": "correct | failed | uncertain",
      "evidence": "<brief frame-indexed facts>"
    }
  ],
  "track_and_turning_audit": {
    "arc_verdict": "valid | invalid | uncertain",
    "turning_verdict": "natural | defective | uncertain",
    "evidence": "<brief>"
  },
  "compound_audit": {
    "wrong_response": "yes | no | uncertain",
    "wrong_evidence": "<brief independent evidence or none>",
    "identity_defect": "yes | no | uncertain",
    "identity_evidence": "<brief separate evidence or none>",
    "independent": "yes | no",
    "cap": "none | 0.35 | 0.20 | 0.10"
  },
  "high_score_allowed": "yes | no",
  "issues": [
    {
      "issue_type": "<allowed value>",
      "description": "<one concise frame-indexed visible fact>",
      "severity": "minor | moderate | severe"
    }
  ],
  "completion_rationale": "<at most 80 words; include applied cap>",
  "scene_completion_score": <number from 0.0 to 1.0>
}
```

Use an empty issues list only when no concrete failure is visible. Output raw JSON only, without markdown fences or conversational text.

**Collision Dynamics (ball-induced domino toppling)**

The candidate video is represented below by {num_frames} ordered source frames sampled at 8 FPS. Frame position k corresponds approximately to k/8 seconds. Treat the images as one chronological video sequence, never as simultaneous scenes.

You are a strict temporal-physics critic evaluating an AI-generated inelastic ball-and-domino collision image-to-video clip. Inspect the complete ordered frame sequence sampled at 8 FPS before scoring.

Expected target behavior: "{caption}"

**Collision Dynamics (ball-induced domino toppling) (continued)**

The caption describes the requested outcome, not what actually happened. Base every conclusion on visible temporal evidence. Return exactly one numeric scene_completion_score from 0.0 to 1.0. The causal collision and domino cascade are the primary signal; clear major motion or object-continuity failures can lower the same score. Do not return separate component scores.

**FIRST-FRAME POLICY**

- The first frame is the official conditioning image. Use it as the reference for ball identity/marking, domino count, per-color multiplicity and chain order, placement, appearance, and camera geometry.
- Do not penalize a mismatch already present in the first frame. Penalize only continuation behavior that first appears or worsens later. A failure to perform a requested later action is still scoreable.

**COLLISION AND CASCADE**

- Track the rolling ball until it reaches and visibly contacts the first domino. Verify that toppling begins only after this causal impact.
- Track every caption-specified domino in chain order. Judge whether the requested number falls sequentially from the first toward the last, in the requested direction, and reaches the requested qualitative terminal state. For a one-domino case, judge only that impact and fall.
- At 8 FPS, adjacent dominoes can first appear tilted in the same sampled frame even when contact propagated between samples. Do not call that simultaneous or non-causal unless a later domino visibly starts before its predecessor, an intermediate piece is skipped, or other positive evidence shows the causal link is broken.
- Penalize failure to reach/strike the first domino, spontaneous toppling before contact, wrong fall direction/order, skipped dominoes, too few/many affected dominoes, a stalled cascade, grossly wrong ball response, pass-through, sustained interpenetration, or a clearly delayed non-causal response.
- Completing all dominoes eventually does not erase a visible skipped, stalled, premature, or wrong-order propagation.
- Caption numbers such as exact lean angle and exact normalized speed are approximate visual targets. Penalize only a clear qualitative mismatch; do not estimate a small numerical difference from perspective as a material failure.

**MOTION AND BALL-RESPONSE CHECK**

- Track ball displacement across several consecutive intervals before contact, at impact, while the dominoes settle, and late in the clip. Track each domino through continuous tipping poses.
- Smooth deceleration over several intervals is natural. Visible inelastic contact may also legitimately cause a sharp speed loss, a small rebound, or a near-stop. For a slowly approaching ball, response in the same or next sampled interval can be contact-synchronized when there is no complete idle interval with both ball and domino motionless. For a rapidly approaching ball, apply the stricter speed-dependent causal-timing rule below.
- For this physics-quality score, a short post-contact rebound or settling near the impact point can be fully acceptable even if the caption describes a longer deflection or continued roll. Do not report wrong_ball_response merely because the ball travels a shorter distance or stops naturally after causing the requested fall. Reserve that issue for a gross response such as pass-through, unsupported acceleration, sustained high-speed departure, clearly wrong launch direction, or failure to transfer motion.
- A domino can physically impede the ball at contact. If the ball slows or pauses against a tipping, leaning, or fallen domino and later continues as that obstruction tips, slides, settles, or clears, classify this as natural collision-buffered continuation. The continued motion can be a small rebound or residual roll. It must not be classified as freeze_then_burst merely because several sampled images show little displacement, or because the obstructing domino looks mostly settled before continued motion becomes visible. Such contact-supported slowing and continuation receives no motion penalty.
- Penalize clear jitter, teleportation, pose/trajectory snapping, discontinuous tipping or rotation, flicker, or non-contact motion jumps. Do not infer a defect from one ambiguous sampled frame.

**MANDATORY SPEED-DEPENDENT CAUSAL-TIMING AUDIT**

- Before assigning a score, compare the ball's displacement over at least three consecutive approach intervals with every interval from arrival at the first domino until that domino starts moving. Use the visible ball diameter as the scale. Treat an approach as rapid when the ball moves roughly half a visible ball diameter or more per sampled interval over multiple approach intervals, or is otherwise plainly much faster than the eventual settling motion. Base this classification on the frames, not the caption's stated speed.
- Explicitly identify in motion_evidence: the last clearly moving ball frame, the first near-stationary ball frame, and the first frame with visible domino pose change. State whether the speed loss is gradual, collision-synchronized, or abrupt and unsupported.
- Two consecutive sampled images with the ball in essentially the same position define one complete stationary interval. For a rapid approach, if the domino remains upright and unchanged throughout that interval and first starts tipping only in a later image, classify severe abrupt_unexplained_stop and a temporally non-causal initiating response. Do not require two consecutive stationary intervals: one complete ball-idle/domino-idle interval is already positive evidence of delayed response after a rapid approach.
- Mere adjacency or apparent touching is not proof that a prior speed collapse was collision-caused. For a rapid approach, the first interval in which ball displacement collapses must overlap visible domino pose change to count as collision-synchronized. A domino that first changes only after a complete stationary interval is delayed even if it moves in the immediately following sampled image.
- Preserve sampling tolerance for a genuinely slow approach: when the ball was moving less than roughly half its diameter per sampled interval, contact depth is ambiguous, and no complete ball-idle/domino-idle interval is visible, the domino may first tilt in the next sampled image without penalty. A slow ball stopping or slightly rebounding as that response begins is natural inelastic settling.
- If a pre-response stationary gap is followed by a delayed domino fall, treat the initiating collision as temporally non-causal. This rule remains strict before the first domino has visibly responded.
- After collision-synchronized domino response begins, pause duration alone is not evidence of freeze_then_burst. Do not report it when contact geometry explains that a domino temporarily blocks the ball and the ball later resumes continuously as the obstruction settles or clears. Report freeze_then_burst only with positive evidence that the restart itself is unsupported: for example, the ball jumps to a new position, changes from stationary to clearly high speed in one interval without a new contact or force, adopts an impossible direction, or resumes away from any plausible obstruction. A gradual collision-buffered continuation is natural, not an unsupported acceleration.
- Use the abrupt stop as the primary issue rather than duplicating the same evidence under several issue types. A gradual slowdown or a speed loss whose first interval overlaps domino pose change is not an abrupt-stop failure.

**Collision Dynamics (ball-induced domino toppling) (continued)**

**MANDATORY FRAME-BY-FRAME OBJECT-INVENTORY AUDIT**

- Perform this as a distinct pass before assigning the score. First write down the exact first-frame inventory: one ball plus the total domino count, the per-color multiplicity, and the chain order. Then count and track those bodies through EVERY sampled image from first contact until the pile settles, not only the final image.
- Pay special attention to the peak-overlap frames and the first clear settled frames. Count distinct rigid rectangular bodies using their separately visible long edges, ends, colors, positions, and orientations. For repeated colors, explicitly verify that the visible multiplicity still equals the first-frame multiplicity. A plausible-looking pile is not proof that its inventory is correct.
- A single domino can be partially occluded, so do not double-count two visible portions that remain geometrically compatible with one rigid rectangle. But do not merge two separately oriented or spatially distinct same-color slabs merely because they overlap. If a new slab emerges while every original is still accounted for and the extra body remains visible in at least two sampled images or in the settled pile, that is positive multi-frame evidence of duplication.
- In object_continuity_evidence, explicitly report the first-frame per-color counts and the peak/settled per-color counts that support preserved, violation, or uncertain. Never claim preserved by only restating the expected inventory without accounting for the visible bodies during and after impact.
- Set object_continuity_status to violation only with positive multi-frame evidence that an object gains an extra copy, splits, permanently disappears, fuses into another geometry, or swaps color/identity after the first frame.
- Falling pieces may overlap or temporarily occlude each other. Do not call an overlap, pile, one-frame absence, or natural frame-boundary exit a violation without positive temporal evidence.

**CORE-FIRST ONE-SCORE CONTRACT**

- First place the clip into a band from the initiating contact and causal cascade. Then lower it for independently visible major ball-motion, continuity, identity, or scene-stability failures.
- Do not let a subtle endpoint angle, uncertain image-plane direction, or approximate speed outweigh a clear difference in collision/cascade quality.
- Conversely, a clean final pile does not rescue a broken causal sequence, and a correct cascade with a clear duplication/teleport is not fully successful.
- 0.90-1.00: convincing causal contact and complete correct cascade, with coherent required trajectories and identities; at most negligible defects.
- 0.70-0.89: central event succeeds with only minor visible defects.
- 0.45-0.69: substantial event remains valid but one material failure is clear.
- 0.15-0.44: severe central failure or multiple material failures.
- 0.00-0.14: catastrophic failure dominates or the event is untrackable.
- Missing/non-causal initiating contact caps the score at 0.30.
- A clearly stalled, skipped, incomplete, wrong-order, or wrong-direction primary cascade caps it at 0.45.
- A gross post-impact ball response such as unsupported acceleration, sustained high-speed departure, or clearly wrong launch direction caps it at 0.55. A natural contact-synchronized stop or small rebound has no such cap.
- Clear object duplication, fusion, permanent disappearance, or identity swap caps it at 0.35; multiple identity failures cap it at 0.15.
- A clear rapid-to-stationary abrupt stop before the first domino visibly responds caps the score at 0.30, even if the cascade later completes.
- After a rapid approach, one complete stationary interval in which the ball and domino both remain unchanged, followed only later by the domino fall: cap the score at 0.20. Two consecutive same-position ball images are sufficient to establish that pre-response interval; do not require two stationary intervals. A later freeze-then-burst caps the score at 0.20 only when positive evidence shows an unsupported high-speed restart, jump, impossible direction, or other discontinuity. Never apply this cap to natural collision-buffered continuation while a domino obstructs and then releases the ball.
- A major teleport caps the score at 0.30; if it prevents contact, apply the lower supported cap.
- Apply the lowest supported cap, but invoke no cap without positive temporal evidence. Do not impose a cap solely for approximate angle/speed mismatch or same-sampled-frame adjacent tipping.

Use only these issue_type values and emit each supported defect once:

- missing_or_noncausal_first_contact
- wrong_fall_order_or_direction
- wrong_fallen_count_or_terminal
- stalled_or_skipped_cascade
- wrong_ball_response
- pass_through_or_interpenetration
- jitter_or_snapping
- teleport_or_jump
- abrupt_unexplained_stop
- freeze_then_burst
- identity_or_count_failure
- camera_or_background_instability

Return only one valid JSON object matching this exact schema:

```
{
  "issues":  [
    {
      "issue_type":  "<one allowed value above>",
      "category":  "Collision Dynamics | Motion/Speed Continuity | Object
Conservation/Identity | Scene Stability",
```

## Collision Dynamics (ball-induced domino toppling) (continued)

```
      "description":  "<specific visible evidence, affected object, and approximate
interval>",
      "severity":  "minor | moderate | severe"
    }
  ],
  "cascade_status":  "complete_causal | incomplete_or_stalled | noncausal | uncertain",
  "cascade_evidence":  "<chronological contact, domino order/count/direction, and terminal
evidence>",
  "object_continuity_status":  "preserved | violation | uncertain",
  "object_continuity_evidence":  "<first-frame inventory and collision-interval
count/identity evidence>",
  "abrupt_unexplained_stop_status":  "present | absent | uncertain | not_applicable",
  "motion_evidence":  "<pre-contact, impact, settling, and late-trajectory evidence>",
  "completion_rationale":  "<brief core-first rationale and any applied cap>",
  "scene_completion_score":  <number from 0.0 to 1.0>
}
```

Use an empty issues list only when no concrete failure is visible. Output raw JSON only, without markdown fences or conversational text.

## Wave Propagation

**VIDEO INPUT**

This request contains exactly one ordered video sequence. Every FRAME_ID label applies only to the image immediately following it. The image sequence is delimited by BEGIN VIDEO and END VIDEO. Each frame is introduced by "FRAME_ID F001. Next ordered sample from this same video.", with F001, F002, and subsequent labels increasing in temporal order.

You are a temporal-physics critic evaluating one AI-generated image-to-video clip of wave propagation along a rope. Inspect the complete ordered frame sequence sampled at 8 FPS before scoring.

**INPUT CONTRACT**

- The requested initial scene and behavior are: "{caption}"
- The caption is the target specification, not evidence that the behavior happened.
- The video images are labeled FRAME_ID F001, F002, and so on in true temporal order. Cite these labels for every material issue and evidence claim. Return one holistic scene_completion_score from 0.0 to 1.0. Do not create, calculate, or return component scores. The only JSON number must be scene_completion_score.

**PRIMARY STANDARD**

- The dominant criterion is a causal, temporally smooth, continuous, sustained traveling sine-like wave on an intact rope. A useful clip shows controller motion producing a nearby rope response, then a recognizable alternating crest-and-trough or equivalent lateral phase pattern advancing through neighboring rope sections through early, middle, and late intervals. Rope integrity is equally central: the same physical rope must remain attached, continuous, and topologically consistent enough that propagation can be traced. Visual polish, prompt wording, a plausible static scene, a single broad arc, or generic whole-rope swaying cannot compensate for failure to form the requested wave, jerky or incoherent motion, disappearing propagation, rope splitting, extra rope legs, detachment, or unreadable topology.
- Task success requires positive evidence of a recognizable sine-like spatial waveform, not merely controller movement or rope displacement. At one or more intervals, identify a crest with an adjacent trough, two alternating bends, or an equivalent perspective-consistent phase pattern spanning a material part of the rope, then show that pattern traveling away from the source. A straight rope that lifts, one broad catenary arc, rigid-span motion, whole-rope translation, source-local flapping, or irregular non-wave wobble is not a successful traveling sine wave.
- Requested motion family, phase, and direction still matter, but they are secondary to physical wave quality and rope integrity. A strong causal sustained wave on an intact rope should receive substantial credit even when the requested family, phase relation, circular direction, or controller direction is imperfect. Treat such prompt-identity mistakes as modest-to-material penalties according to how much they change the visible behavior; do not impose a hard low cap solely because the requested family or direction is wrong.
- Use this internal rubric as judgment guidance while returning only the one score:
- 35 percent: recognizable sine-wave formation, source-to-wave causality, local rope response, and plausible travel delay.
- 30 percent: frame-to-frame motion coherence plus smooth, continuous, sustained propagation through early, middle, and late intervals.
- 20 percent: rope integrity, topology, attachment, and absence of extra or missing rope legs.
- 10 percent: requested motion-family, phase, direction, amplitude, and frequency identity.
- 5 percent: anchor/floor plausibility, visibility, camera stability, and rendering artifacts.
- These percentages guide ordinary cases, but the evidence-gated caps and the explicit approximately 0.30 low-frequency deduction below override nominal weighting when their conditions are met.
- Missing or inadequate sine-wave formation, temporal incoherence, propagation discontinuity, absence/weakness of the wave, and rope-integrity failures below activate evidence-gated caps and cannot be averaged away by visual quality.

**REQUIRED CHRONOLOGICAL AUDIT**

1. At F001, inventory the controller or controllers, one physical rope, its visible leg or legs, the fixed anchor, and the floor. Determine whether the rope begins intact and attached, and whether the visible leg count/topology is compatible with the caption.
2. Locate the first visible controller/human hands displacement. Look for a local rope response at that controller before accepting any later wave as causal. If the source moves but the rope barely responds, flag an absent or weak wave.
3. Before granting task success, find positive evidence of a sine-like spatial waveform: a crest adjacent to a trough, two alternating bends, or an equivalent visible phase pattern extending beyond the source-local segment. Cite frames showing the waveform itself; do not infer it from hand motion, one broad bend, or the caption.

**Wave Propagation (continued)**

4. Select a crest, trough, or phase feature from that waveform and trace its position over three or more labeled frames from source toward anchor. Judge travel delay, direction, speed consistency, amplitude persistence, and whether the feature evolves continuously rather than appearing at unrelated positions.
5. Inspect consecutive sampled frames for temporal coherence and motion smoothness. Follow the controller, source-near segment, mid-span segment, and far segment; flag abrupt pose replacement, rope-shape popping, stutter, unexplained stop-start motion, sudden speed or direction changes, or nonphysical snapping even when coarse early-to-late positions look plausible.
6. Audit early, middle, and late intervals. Look specifically for a wave that fades to an almost static rope, freezes and then jumps, vanishes and restarts, changes phase without cause, snaps the whole rope simultaneously, appears independently in the middle of the rope, or dies out while the controller continues moving.
7. Audit topology throughout the sequence with a rope-leg ledger rather than checking only attachment or a few clean frames: establish the F001 expected leg inventory, inspect the maximum-complexity frames in every early, middle, and late interval, and count textured physical rope centerlines at fixed transverse gates immediately beyond the handles, through an unobstructed mid-span, and near the anchor or stable crop boundary; look specifically for disappearing rope, disconnected splits, one leg becoming multiple extra legs, a duplicate/ghost strand, a lost required leg, detachment, or topology switching.
8. Estimate frequency from the full ordered sequence, never one frame: use the source endpoint or controller to count complete high-low-high, left-right-left, or circular cycles, derive the observable duration from FRAME_ID count and 8 FPS, and compare the observed cycle rate with the captioned target.
9. Identify which of the seven motion families is visible and compare its phase, direction, and topology with the caption: single_vertical, single_lateral, double_alternating, double_synchronous, double_inward_circles, double_outward_circles, or double_lateral. Penalize wrong identity, but keep the penalty separate from sine-wave formation, temporal coherence, causal continuity, and rope integrity.
10. Inspect attachment, floor penetration, the fixed anchor, and visible reflection from the absorbed anchor throughout the full sequence.
11. Apply only evidence-supported caps and the explicit low-frequency deduction, then select a direct score using the anchors below. Do not stack penalties for the same underlying defect.

**CAUSAL SINE-WAVE FORMATION AND TRAVEL**

- The controller is the sole intended source. Accept a short causal delay between controller motion and a nearby rope response, followed by a later response farther toward the anchor. Do not require exact pixel motion or exact recovery of the caption's numeric wave speed.
- A qualifying waveform visibly alternates displacement along the rope: at least one crest and adjacent trough, two opposite-curvature bends, or an equivalent perspective-consistent phase pattern must occupy a material span rather than only the controller-adjacent tip. It need not be a mathematically perfect sinusoid, but it must be recognizably sine-like rather than a straight lift, one broad arc, rigid-span sway, whole-rope translation, or random wobble.
- A coherent crest, trough, or phase feature from the qualifying waveform should advance through neighboring rope sections. Small sampling jumps at 8 FPS, perspective compression, motion blur, and brief occlusion are not discontinuities by themselves.
- Sustained quality matters. A clip with no recognizable sine-like waveform, only a weak source twitch, a completely absent wave, a single broad lift, or a nearly static rope after onset must set waveform_quality to no_sine_like_wave and must not score above 0.10. Preserve the separate causal observation: controller motion may cause a broad rope displacement while still failing the required sine-wave task.
- Reserve waveform_quality weak_or_deformed for a real alternating crest-and-trough or phase pattern that exists but is shallow, distorted, brief, or localized; reserve source_wave_causality weak_or_delayed for a real traveling waveform whose origin or delay is degraded but still visibly traceable through the rope.
- If the controller moves and the rope only twitches, bulges, or changes shape beside the source while no crest, trough, or phase feature subsequently enters the mid-span, the wave has not propagated. Set source_wave_causality to noncausal_or_absent, add a severe Source-to-Wave Causality issue, and cap the score at 0.10; do not downgrade this failure to moderate weak_or_delayed merely because a source-local deformation is visible.
- A wave appears independently in the middle of the rope when positive multi-frame evidence shows a new bend or oscillation arise far from the source before intervening rope sections respond. This is a central causal failure.
- Penalize apparent full-rope rigid snapping, simultaneous unrelated oscillation, sustained wrong-way travel toward the controller, freeze-then-jump motion, repeated phase resets, or a crest that disappears at one location and appears far away without plausible propagation through the intervening rope.
- Compare early, middle, and late cycles. One clean cycle does not excuse later freezing, disappearance, decay to a static rope while the controller continues, or spontaneous restart. Smooth physical damping is allowed when consistent with the caption and still traceable as propagation.

**TEMPORAL MOTION COHERENCE AND SMOOTHNESS**

- Judge adjacent-frame evolution, not only whether a similar pose appears at widely separated times. Smooth coherent motion means the controller trajectory, rope curvature, crest position, trough position, and contact state evolve through plausible intermediate states with locally consistent direction and speed.
- Mark mostly_smooth only for isolated small timing or shape irregularities that do not prevent tracking the waveform. Mark materially_jerky when any affirmatively verified material jerk or pop, repeated stutter, abrupt velocity change, implausible pose replacement, stop-start behavior, or unstable curvature materially degrades an otherwise traceable wave; these are critical motion-coherence failures and the clip must not score above 0.10.
- Mark severely_discontinuous when motion repeatedly skips required intermediate states, freezes and resumes elsewhere, teleports a waveform, abruptly resets phase, or replaces the rope shape so strongly that continuous motion cannot be followed; this is a critical failure and must not score above 0.10.
- Fast motion is not automatically jerky, and 8-FPS sampling can make a fast crest move several pixels between frames. Require repeated multi-frame evidence of inconsistent evolution, not one blurred or widely spaced sample, before applying a smoothness penalty or cap.

**FREQUENCY FIDELITY**

- Judge frequency from repeated complete source cycles across the full observable duration. A complete vertical cycle is high-low-high or low-high-low, a lateral cycle is left-right-left or right-left-right, and a circular cycle is one full return around the trajectory; do not count every crest, trough, direction reversal, or sampled frame as a complete cycle.

**Wave Propagation (continued)**

- Set frequency_fidelity to materially_too_low only when ordered-frame evidence shows that the observed cycle rate is clearly far below the captioned target, generally about 60 percent or less of the target or only a few broad cycles where the target requires many. Cite the counted cycle landmarks and observable interval in frequency_evidence without adding any numeric JSON field.
- For materially_too_low, first judge the score warranted by all other evidence and then subtract approximately 0.30 exactly once; the final score must not exceed 0.70. Add at least one moderate or severe Motion/Phase Identity issue. Do not double-count the same low frequency as weak propagation unless it also prevents a sustained traveling waveform.
- A small count uncertainty, perspective compression, partial endpoint occlusion, or slight rate mismatch is slightly_mismatched and receives a smaller proportional deduction. Use untrackable rather than inventing a cycle count when the source endpoint cannot be followed reliably.
- If the frequency is so low that no recognizable sustained sine-like traveling waveform forms, apply the stricter no_sine_like_wave or absent-wave 0.10 cap instead of the approximately 0.30 deduction.

**MOTION-FAMILY AND PHASE IDENTITY**

- single_vertical: one rope leg responds to the captioned vertical source motion.
- single_lateral: one rope leg responds to the captioned lateral source motion.
- double_alternating: the two source-side legs exhibit opposing alternating phase, rather than moving together throughout.
- double_synchronous: both source-side legs move together in the same phase, rather than consistently opposing each other.
- double_inward_circles and double_outward_circles: verify the captioned circular direction from temporal trajectories, not one still image. Perspective can hide part of a circle, so require multi-frame evidence before calling the inward/outward identity wrong.
- double_lateral: both source-side legs follow the captioned lateral relationship.
- The caption controls exact direction, amplitude, frequency, phase, and which controller moves. Judge these properties qualitatively across time. Do not infer frequency from one frame or penalize small perspective-dependent amplitude differences.
- Wrong requested family, phase relation, or circular/lateral direction is a prompt-identity defect, not by itself proof that wave propagation failed. When the wave is causal, continuous, sustained, and on an intact rope, usually keep the score in a useful range and subtract only the amount warranted by the visible mismatch.

**ROPE INTEGRITY, TOPOLOGY, AND CONTACT**

- A dark texture band is not a rope break. Neither are a specular highlight, shadow, motion blur, self-crossing, foreshortening, or brief occlusion. Confirm a break only when a persistent spatial gap, free severed end, detachment, or incompatible topology is positively visible across multiple ordered frames.
- Establish the expected physical-leg count from F001 and the caption, then seek the maximum clearly visible count rather than proving integrity from only a convenient clean interval. At each of the source-near, open mid-span, and anchor-or-crop-boundary gates, compare early, middle, and late frames and report whether the same expected inventory crosses the gate.
- A normal X-shaped self-crossing has exactly two textured centerlines entering and two leaving, and temporary overlap can hide a leg without creating one. By contrast, a Y- or T-like fork, a third full-width braided centerline crossing a gate, an extra parallel track, a duplicate loop with no controller or anchor source, one handle feeding multiple long strands, or one or two intended legs becoming three or more is positive evidence of hallucinated rope multiplication.
- If swinging causes an additional physical rope or leg to appear, treat it as topology corruption even when no single sampled frame exposes the exact branch point, even when the added strand later merges, crosses, disappears, or exits the crop, and even when later clean frames again show the expected count. A clear count mismatch across multiple ordered frames or neighboring gates is sufficient; do not require a persistent Y-junction.
- When the anchor is outside the image, use a stable frame boundary as the far gate: two intended legs becoming three textured rope bodies before crossing that same boundary is multiplication, not ordinary crossing. When the anchor is visible, trace exactly the expected legs into the common turn and flag an extra arrival, small uncaused loop, free branch, or strand that terminates or appears beside the anchor.
- Separate physical rope from shadow before counting. A physical strand has rope-like width and braided texture, participates in occlusion, and can be traced across neighboring frames; a floor shadow is softer or lower contrast, remains projected on the floor plane, follows a physical rope, and must not be counted as another leg. If this distinction cannot be resolved, use ambiguous_overlap_or_shadow rather than intact or extra_leg_confirmed.
- Disappearing rope sections, splitting into disconnected strands, lost required legs, or a topology that changes incompatibly between early, middle, and late frames are central failures when supported by multi-frame evidence. Score them by how much they prevent tracing a single physical rope.
- Check that the source end stays attached to its controller and the far anchor stays fixed. Flag stretch, detachment, topology switching, floor penetration, or a visible reflected wave from a captioned absorbing anchor only when the ordered evidence is clear.
- Camera motion or a changing crop can alter apparent rope position. Separate camera/background instability from rope dynamics and lower the physics score only to the extent that it prevents reliable observation.

**EVIDENCE-GATED CAPS**

- A cap is a maximum, never an assigned score and never a way to raise a weaker result. Use the lowest cap supported by affirmative multi-frame evidence.
- 0.10 maximum: no recognizable sine-like traveling waveform; any verified material propagation or temporal-motion discontinuity; any verified material jerk or pop, repeated stutter, stop-start motion, or abrupt pose/rope-shape replacement; no causal traveling wave; only a weak source twitch, single broad arc, rigid-span sway, whole-rope translation, or source-local motion without traceable sine-wave travel; a completely absent wave; a wave repeatedly originating away from the controller; or a material teleport, reset, disappearance, freeze-then-jump, or phase break.
- 0.20 maximum: the rope visibly breaks, detaches from the controller or anchor, disappears for a material interval, splits into disconnected strands, grows extra rope legs/ghost ropes, loses required legs, or changes to an incompatible topology that makes propagation unreadable; later recovery to the expected leg count does not remove this cap.
- 0.45 maximum: a recognizable sine-like wave is present but too weak, deformed, brief, intermittent, or localized to count as sustained propagation through the rope, with no verified jerky, popping, or other critical temporal-motion failure.

**Wave Propagation (continued)**

- 0.50 maximum: a major floor/anchor violation occurs while the rest of the wave remains partly readable and has no material propagation discontinuity.
- 0.70 maximum: the traveling waveform is otherwise readable but the observed source-cycle frequency is materially too low; also subtract approximately 0.30 from the score warranted by the remaining evidence, then apply any stricter cap.
- No hard cap may follow from one ambiguous frame, sampling sparsity, blur, occlusion, perspective, a merely possible defect, or only a wrong requested motion family/phase/direction on an otherwise causal sustained intact wave.

**HIGH-SCORE PROTECTION**

- A recognizable sine-like crest-and-trough or phase pattern that is source-caused, temporally smooth, and traceable progressively along an intact rope through early, middle, and late intervals is affirmative high-score evidence.
- Allow an otherwise coherent waveform to remain high when the only concerns are a minor texture artifact, a brief occlusion, small camera motion, an imprecise amplitude, or a wrong requested direction/family detail that does not undermine causal sustained propagation. Do not demand an analytically perfect sinusoid, exact wave speed, or exact parameter recovery from pixels, but do require a visibly recognizable sine-like waveform and smooth temporal evolution.

**DIRECT ONE-SCORE ANCHORS**

- 0.90-1.00: a clean, causal, temporally smooth, continuous, sustained traveling sine-like wave; intact rope, stable attachment, plausible anchor/floor behavior, and correct family/phase.
- 0.75-0.89: strong useful sustained sine-like propagation on an intact rope with one modest local timing, amplitude, prompt-identity, camera, or visual imperfection and no material smoothness defect.
- 0.51-0.74: a recognizable, mostly smooth traveling sine-like wave on a mostly intact readable rope with no stricter capped waveform, causal, continuity, or temporal-motion failure, but with materially low frequency after the approximately 0.30 deduction, limited visibility, a moderate noncritical artifact, or a substantially wrong requested family/phase/direction.
- 0.30-0.50: a recognizable sine-like waveform exists but propagation is weak, deformed, short-lived, intermittent, or compromised by a major anchor/floor problem, with no verified jerky or popping motion; respect the 0.45 weak-waveform cap and the 0.50 floor/anchor cap whenever they apply.
- 0.11-0.29: a severe topology, attachment, or visibility failure prevents a useful episode, but none of the stricter 0.10 waveform, causal, propagation, or temporal-motion conditions is verified. Respect the lower 0.20 rope-integrity cap whenever it applies.
- 0.05-0.10: a critical failure is verified, including no recognizable sine-like traveling waveform, but some attempted source motion or rope behavior remains readable.
- 0.00-0.04: no meaningful wave-propagation episode can be evaluated, including a completely absent wave with no readable traveling response or waveform.
- Use the full continuous scale. A 0.05 difference is small, 0.10 is material, and 0.20 or more is decisive. Do not average issue severities, add independent subscores, cluster at anchor boundaries, or let visual cleanliness compensate for a missing sine-like waveform, materially incoherent motion, verified absent/weak propagation, or central rope-topology failure.

**OUTPUT CONTRACT**

- Use Sine-Wave Formation for missing, broad-arc-only, rigid-span, nonalternating, weak, or deformed waveform failures; Source-to-Wave Causality for source onset, travel direction, causal-delay, absent-wave, or source-only weak-twitch failures; Wave Propagation Continuity for spatial travel, freezes, jumps, teleportation, resets, disappearing waves, and weak/short-lived propagation; Temporal Motion Coherence for stutter, popping, abrupt pose or rope-shape replacement, implausible velocity changes, and incoherent frame-to-frame motion; Motion/Phase Identity for wrong family, direction, phase, amplitude, or frequency behavior; Rope Integrity/Topology for breaks, detachment, disappearing rope sections, splitting, stretch, extra/lost rope legs, or topology corruption; Floor/Anchor Physics for penetration, anchor motion, or reflection; Scene Stability/Visibility for camera/background problems; and Prompt Alignment only for a remaining caption mismatch not already covered by a physics issue.
- Every noncausal_or_absent verdict must include at least one severe Source-to-Wave Causality issue with FRAME_ID evidence. Use weak_or_delayed with a minor or moderate issue only when a real crest, trough, or phase feature visibly leaves the source region and enters successive rope sections but its causal timing or delay is degraded.
- Every materially_too_low frequency verdict must include at least one moderate or severe Motion/Phase Identity issue with FRAME_ID evidence and must reflect the approximately 0.30 deduction in the final score before any stricter cap.
- Every extra_leg_confirmed or missing_leg_confirmed verdict must set rope_integrity to topology_corrupted or broken_or_detached as applicable, include at least one severe Rope Integrity/Topology issue with FRAME_ID evidence, and respect the 0.20 maximum. Every broken_or_detached or topology_corrupted rope_integrity verdict has the same severe-issue and 0.20-cap requirements.
- Return only one valid JSON object matching this exact schema. The only JSON number must be scene_completion_score:

```
{
  "issues": [
    {
      "category": "Source-to-Wave Causality | Sine-Wave Formation | Wave Propagation
Continuity | Temporal Motion Coherence | Motion/Phase Identity | Rope Integrity/Topology |
Floor/Anchor Physics | Scene Stability/Visibility | Prompt Alignment",
      "description": "<Concrete chronological evidence with FRAME_ID labels and affected
rope section>",
      "severity": "minor | moderate | severe"
    }
  ],
  "observed_motion_family": "single_vertical | single_lateral | double_alternating |
double_synchronous | double_inward_circles | double_outward_circles | double_lateral |
ambiguous | untrackable",
  "waveform_quality": "clear_sine_like | recognizable_sine_like | weak_or_deformed |
no_sine_like_wave | untrackable",
  "source_wave_causality": "strong_causal | mostly_causal | weak_or_delayed |
noncausal_or_absent | untrackable",
```

**Wave Propagation (continued)**

```
  "propagation_continuity":  "continuous_natural | minor_local_artifact |
material_discontinuity | major_reset_or_teleport | untrackable",
  "temporal_motion_quality":  "smooth_coherent | mostly_smooth | materially_jerky |
severely_discontinuous | untrackable",
  "frequency_fidelity":  "matched_or_close | slightly_mismatched | materially_too_low |
materially_too_high | untrackable",
  "rope_integrity":  "intact | ambiguous_occlusion | broken_or_detached |
topology_corrupted | untrackable",
  "rope_leg_inventory":  "consistent | ambiguous_overlap_or_shadow | extra_leg_confirmed |
missing_leg_confirmed | untrackable",
  "motion_phase_identity":  "correct | mostly_correct | wrong_phase_or_direction |
wrong_motion_family | untrackable",
  "anchor_floor_behavior":  "plausible | minor_artifact | major_floor_or_anchor_violation |
untrackable",
  "source_wave_evidence":  "<FRAME_ID evidence for controller onset and nearby rope
response>",
  "waveform_evidence":  "<FRAME_ID evidence for or against a recognizable sine-like
crest-and-trough or equivalent phase pattern>",
  "propagation_evidence":  "<FRAME_ID evidence tracing a feature through early, middle, and
late motion>",
  "temporal_motion_evidence":  "<FRAME_ID evidence for smooth or incoherent adjacent-frame
evolution>",
  "frequency_evidence":  "<FRAME_ID cycle landmarks and full-interval evidence comparing
observed and captioned frequency>",
  "rope_leg_inventory_evidence":  "<FRAME_ID ledger comparing expected and maximum visible
physical rope legs at source, mid-span, and anchor or crop-boundary gates>",
  "topology_contact_evidence":  "<FRAME_ID evidence for integrity, attachment, topology,
anchor, and floor>",
  "completion_rationale":  "<Why the evidence and any applicable cap justify the one
score>",
  "scene_completion_score":  <number from 0.0 to 1.0>
}
```

- Use an empty issues list only when no concrete failure is visible. Output raw JSON only, without markdown fences or conversational text.

# B Analysis

## B.1 Alignment of VLM judge and human evaluation

We compare the frozen VLM judgments with human ratings along two distinct axes: whether the evaluators agree on which model produces the better video, and whether they assign similar score differences to the same scene pairs. The second question concerns the magnitude of the *model score gap*, rather than exact equality between human and VLM scores.

**Evaluation protocol.** We conduct human evaluation of 190 photorealistic videos, organized into 95 scene-level pairs of real-only-SFT and Sim-$\infty$ outputs: 30 OP pairs, 30 CD pairs (10 per collision family), and 35 WP pairs (5 per motion family). Scenes and a shared generation seed for each pair are sampled without consulting model scores. Human review presents both full videos for a scene together, with blinded model identities and randomized positions, allowing direct comparison including ties. Each video receives a separate completion score in $[0, 1]$ with increments of 0.01, using the same scene context and physical criteria as the VLM judge. In contrast, the VLM scores each video independently to avoid potential within-pair presentation-position bias. Its ranking is derived afterward from the two scores, not elicited through a two-video comparison. We compare the human and frozen VLM scores of the *same videos*.

For evaluator $r \in \{\text{human}, \text{VLM}\}$ and scene $i$, define

$$\Delta_i^{(r)} = s_{i,\text{Sim-}\infty}^{(r)} - s_{i,\text{real-only-SFT}}^{(r)}, \qquad d_i^{(r)} = \left|\Delta_i^{(r)}\right|.$$

The sign of $\Delta_i^{(r)}$ determines the ordering; $d_i^{(r)}$ measures how far apart the evaluator scores the two models on that scene.

**Ranking.** Human and VLM rankings agree on 70.00%, 90.00%, and 97.14% of all OP, CD, and WP pairs, respectively, with ties treated as a distinct outcome. When *both* evaluators give unequal scores, ranking agreement is 100.00%, 96.43%, and 97.14%, respectively. Ties correspond to equal scores, without a post-hoc tolerance band. Figure 18 reports the agreement rates and their denominators. All nine OP disagreements involve ties, which may partly reflect the different presentation protocols; none reverses the winner when both evaluators give a strict ordering. CD and WP each contain one opposite ordering.

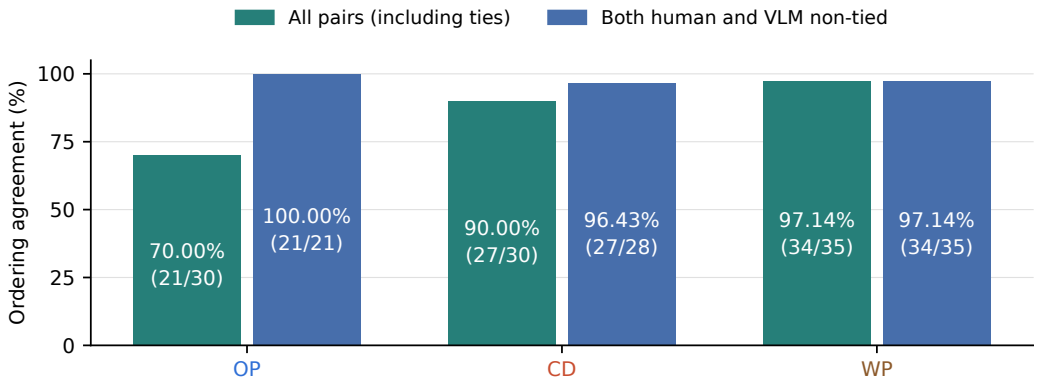


Figure 18: **Strong alignment of human and VLM rankings when both give a strict ordering.** Rankings are derived from independent VLM scores and scores assigned during paired human review. The all-pairs measure treats ties as a separate outcome. The conditional measure includes only pairs for which both evaluators give unequal scores. Each bar reports the agreement percentage and count $(n/N)$.

**Model score-gap magnitude.** We assess two complementary properties. First, Spearman correlation $\rho_s$ between $d_i^{(\text{human})}$ and $d_i^{(\text{VLM})}$ measures whether larger human score gaps also receive larger VLM score gaps. This tests relative magnitude ordering without requiring identical numerical distances. Second, the mean absolute error (MAE) of the gap magnitude,

$$\text{MAE}_{\text{gap}} = \text{mean}_i \left| d_i^{(\text{VLM})} - d_i^{(\text{human})} \right|,$$

measures how closely the two evaluators estimate the model score gap for each scene; lower values indicate closer numerical agreement. Taking the absolute error within each scene prevents over- and underestimation from canceling. We retain equal family weights when computing task-level means.

Figure 19 shows every pair, together with $\rho_s$, $\text{MAE}_{\text{gap}}$, and their 95% confidence intervals estimated from 10,000 bootstrap draws. The magnitude correlations are 0.47, 0.51, and 0.42 for OP, CD, and WP, respectively: positive but moderate correspondence across scenes. The gap-magnitude MAEs are 0.23, 0.17, and 0.21, respectively, quantifying the average per-scene discrepancy in model score-gap magnitude on the $[0, 1]$ scale.

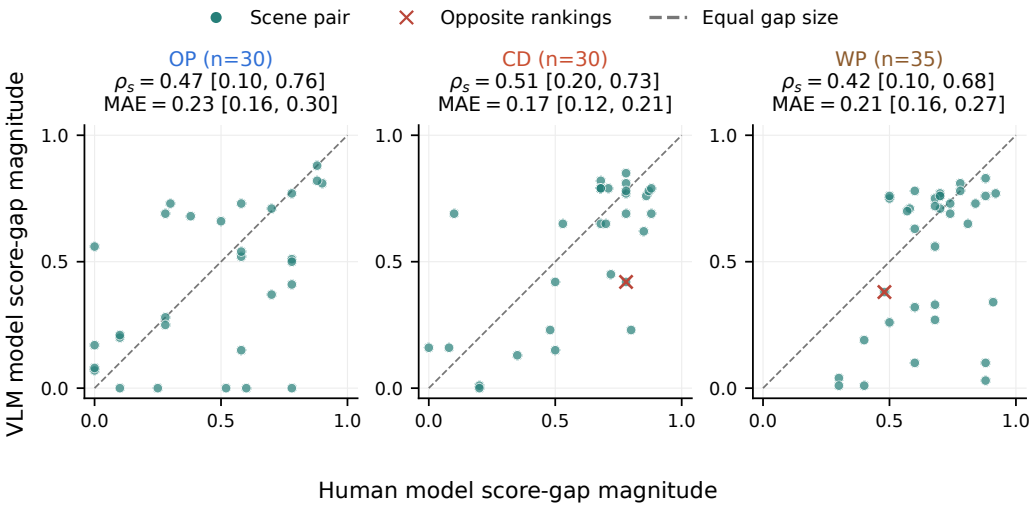


Figure 19: **Moderate alignment of human and VLM model score-gap magnitudes.** Each point is one scene pair, comparing human and VLM absolute score gaps $d_i$. The dashed diagonal denotes equal gap sizes; a point's absolute vertical distance from it is that scene's gap-magnitude error. Red crosses mark opposite model rankings. Panel annotations report magnitude correlation $\rho_s$ and gap-magnitude MAE; brackets contain 95% bootstrap confidence intervals.

**Conclusion.** Overall, the VLM judge agrees closely with human rankings when both evaluators assign a strict ordering, while showing positive but moderate alignment of model score-gap magnitude. The evidence is therefore stronger for ranking agreement than for numerical agreement on the size of the score gap.

### B.2 Failures under the holistic score

Table 3 in Section 3.2 summarizes the leading failure groups on photorealistic videos. Figures 20, 21, and 22 illustrate these failures over time, with task-specific discussion below.

**Object permanence.** Sim-$\infty$ reduces all three diagnosed failure types relative to real-only-SFT; motion problems also remain common in Pretrained and both external models. In Fig. 20, a real-only-SFT clip shows a ball and a marble accelerating *abruptly* after passing a sports bottle, rather than maintaining the requested steady speed. The Sim-$\infty$ clip for the same scene shows smooth, continuous rolling. Frames in each row are sampled at 0.25 s intervals, making the abrupt change in frame-to-frame displacement visible.

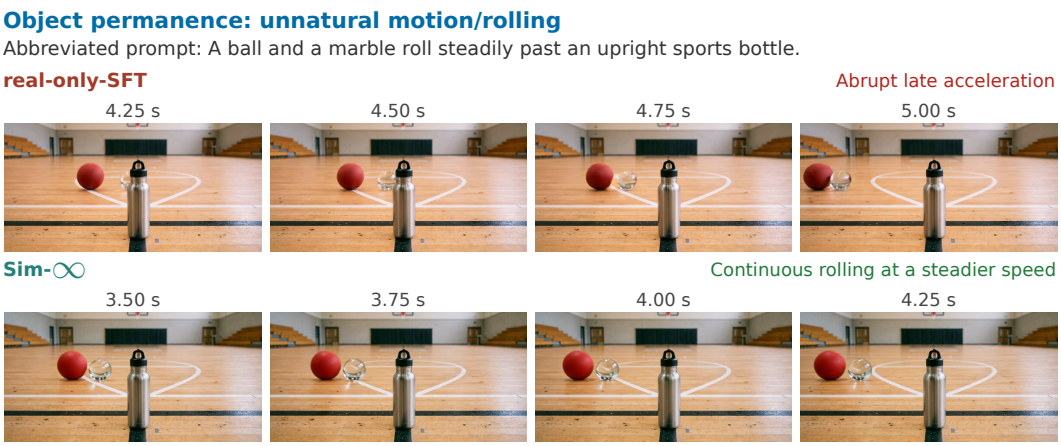


Figure 20: **Object permanence: unnatural motion/rolling.** real-only-SFT shows abrupt late acceleration, whereas Sim-∞ maintains steadier rolling past the bottle. Frames run left to right at 0.25 s intervals, using the same case and crop but event-specific time windows. Red and green descriptions mark failure and correct behavior.

**Collision dynamics.** The overall score rises from 0.30 for real-only-SFT to 0.77 for Sim-∞, with incorrect collision responses falling from 82.9% to 21.7%. The top-1 failure group includes non-causal contact or failed separation in ball impacts, wrong outgoing ends/counts or collective motion in Newton's cradle, and stalled, misordered, or incorrectly terminated cascades in domino toppling. In Fig. 21, MiniMax-H3 relaunches a returning ball from the *same end*, whereas Sim-∞ transfers motion to the opposite end.

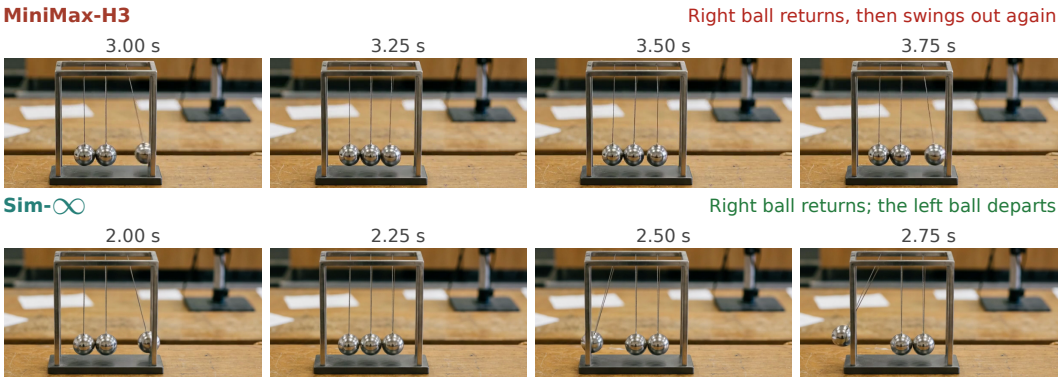


Figure 21: **Collision dynamics: incorrect collision response.** MiniMax-H3 relaunches the returning right ball from the same end; Sim-∞ transfers motion to the opposite end. Frames run left to right at 0.25 s intervals, using the same case and crop but event-specific time windows. Red and green descriptions mark failure and correct behavior.

**Wave propagation.** Sim-∞ and MiniMax-H3 have similar overall scores (0.75 vs. 0.73), but motion-family/phase/direction errors differ markedly (46.3% vs. 84.3%). For the same synchronous-lateral prompt in Fig. 22, MiniMax-H3 instead produces alternating vertical waves, while Sim-∞ follows the requested motion. Thus, generating coherent traveling waves and following the requested motion remain distinct challenges. Sim-∞ exhibits no detected motion-family/phase/direction errors in simulation (0/630), compared with 46.3% (247/534) in photorealistic scenes. This gap may partly reflect the mismatch between controller-driven ropes in simulation and human-driven ropes in photorealistic scenes, as well as ambiguity in prompt descriptions of

different motion families. Future work could explore simulation data that better captures human actuation and caption ablations to assess whether more explicit motion-family descriptions improve control. Despite this gap, Sim-$\infty$ with our simulation data substantially improves both motion control and wave propagation: relative to the pretrained model, motion-control errors decrease by 33.3 percentage points (79.6% to 46.3%), while wave formation/propagation failures fall from 86.0% to 4.7%.

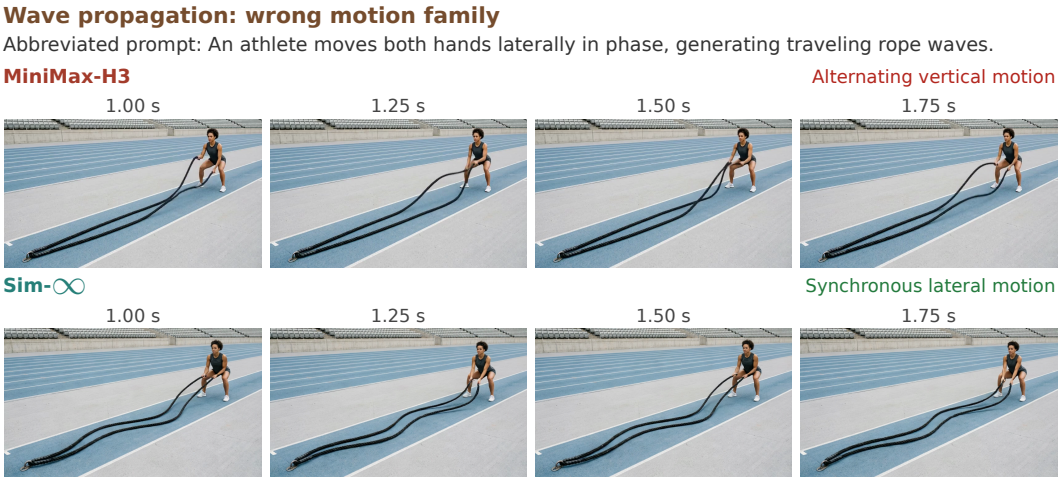


Figure 22: **Wave propagation: motion-family control error.** MiniMax-H3 produces alternating vertical waves instead of the prompted synchronous lateral motion, which Sim-$\infty$ follows. Frames run left to right at 0.25 s intervals, using the same case, crop, and time window. Red and green descriptions mark failure and correct behavior.

### B.3 Statistical reliability of performance gains

We estimate 95% bootstrap confidence intervals (CIs) for the mean score gains of Sim-$\infty$ over real-only-SFT on the full evaluation sets, separately for each of the nine task/domain combinations. For each input scene, we average each model's scores across the three generation seeds and compute the paired difference. We then construct 10,000 bootstrap replicates by sampling with replacement, keeping the two models paired throughout. Each replicate uses the same score aggregation as the main evaluation. Table 11 reports the difference on the original evaluation set and the 2.5th and 97.5th percentiles of the bootstrap differences.

Table 11: Sim-$\infty$ minus real-only-SFT score gains with individual paired 95% bootstrap confidence intervals on OP, CD, and WP benchmarks.

| Domain | OP | CD | WP |
|---|---|---|---|
| Simulation | +0.46 [0.41, 0.51] | +0.46 [0.44, 0.49] | +0.77 [0.74, 0.79] |
| Synthetic | +0.50 [0.46, 0.55] | +0.47 [0.45, 0.50] | +0.74 [0.72, 0.76] |
| Photorealistic | +0.28 [0.24, 0.33] | +0.47 [0.44, 0.49] | +0.50 [0.46, 0.55] |

All nine intervals lie above zero, supporting positive mean score gains on these evaluation sets. The intervals describe evaluation-scene uncertainty conditional on the selected checkpoints and fixed judge, not the range of gains for individual videos.

### B.4 Effect of inference video length on OP-SFT performance

We examine OP-SFT performance across inference lengths, motivated by the uneven representation of video lengths in the training data. OP training uses 73-, 125-, and 189-frame videos, with 125-frame videos accounting for the largest share (Appendix A.2).

We compare six OP-SFT checkpoints from 765 to 4,590 steps with Cosmos3-Super (NVIDIA, 2026) on the same 100 simulation scenes. For each scene, the reference clips use the first 73, 125, or 189 frames of the same simulation, with separate captions describing each clip's events and endpoint. Generation is conditioned on the shared first frame and the corresponding caption, using one seed and 480p output at 24 FPS. We report OP completion scores in $[0, 1]$.

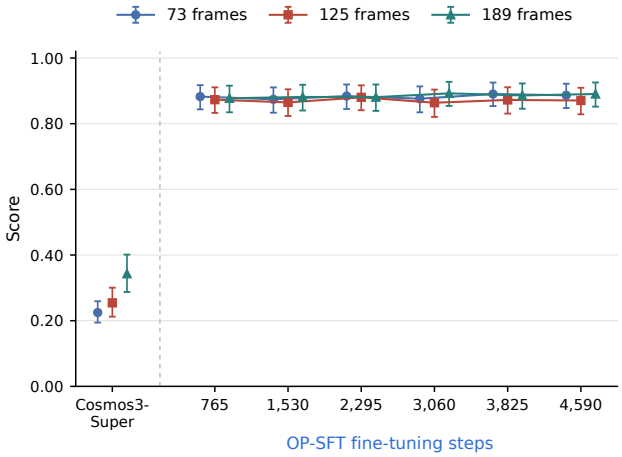


Figure 23: **Effect of inference video length on OP-SFT performance.** Scores on 100 simulation scenes at 73, 125, and 189 frames. Isolated markers show Cosmos3-Super (NVIDIA, 2026); connected markers show six OP-SFT checkpoints. Error bars are pointwise 95% confidence intervals from 10,000 scene-level bootstrap resamples.

Across all checkpoints and inference lengths, mean OP completion scores range from 0.86 to 0.89, and the largest within-checkpoint difference is below 0.03 (Fig. 23). Every OP-SFT checkpoint outperforms Cosmos3-Super at all three matched lengths. These results indicate limited sensitivity to inference length under the mixed-length OP training recipe.

### B.5 Generalization of object permanence from equal-sized to differently sized balls

In our OP simulation training set, all balls within a scene have the same radius. We examine whether the model maintains object permanence when balls of different sizes appear together at inference. Figure 24 shows two photorealistic examples in which the balls pass behind an occluder and reappear with their identities and relative sizes preserved. These examples provide qualitative evidence that object-permanence behavior extends to size configurations absent from the OP simulation training set.

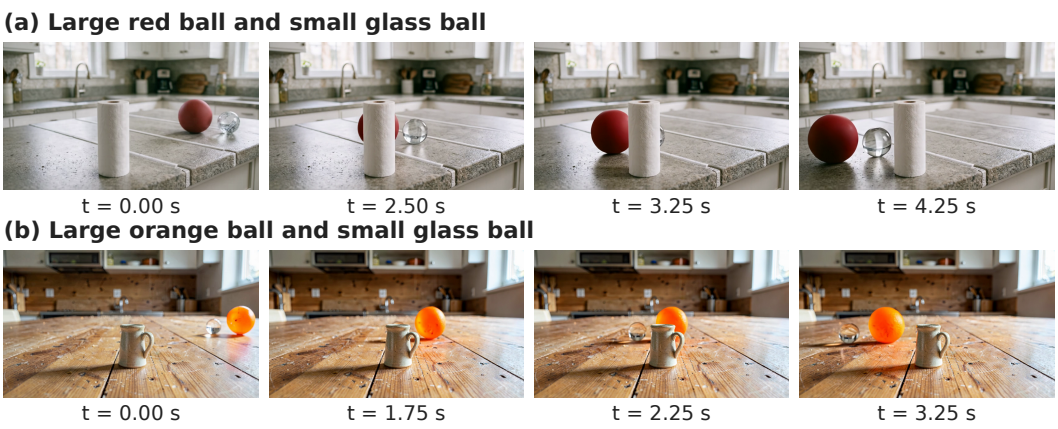


Figure 24: **Generalization of object permanence from equal-sized to differently sized balls.** A large colored ball and a smaller glass ball pass behind (a) a paper roll or (b) a mug and reappear with their identities and relative sizes preserved.

# C MORE QUALITATIVE RESULTS

Figures 25–33 present nine cross-model comparisons in photorealistic scenes, ordered by object permanence, collision dynamics, and wave propagation. Within each figure, all model rows share the displayed timestamps, and the prompt is abbreviated.

## C.1 OBJECT PERMANENCE

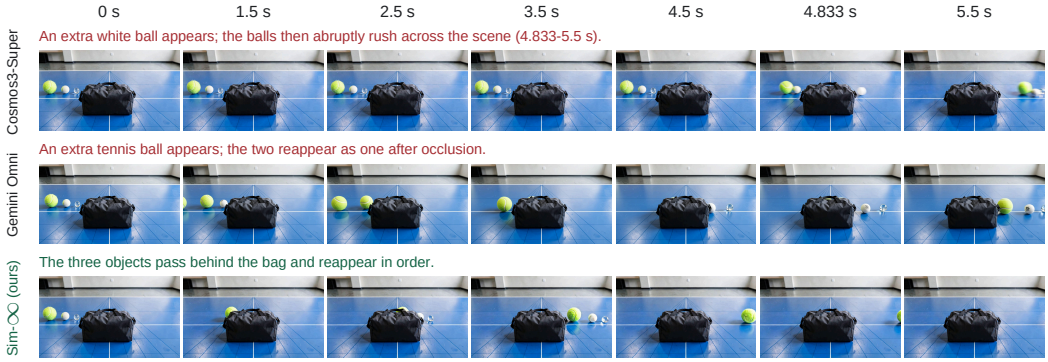


Figure 25: **Object permanence for three balls behind a sports bag.** A tennis ball, ping-pong ball, and marble are prompted to pass behind a stationary bag at equal speed. Sim-$\infty$ preserves the three objects and their order. Cosmos3-Super introduces an extra white ball and an abrupt burst of motion between 4.833 and 5.5 s. Gemini Omni introduces a second tennis ball, but the two reappear as one after occlusion.

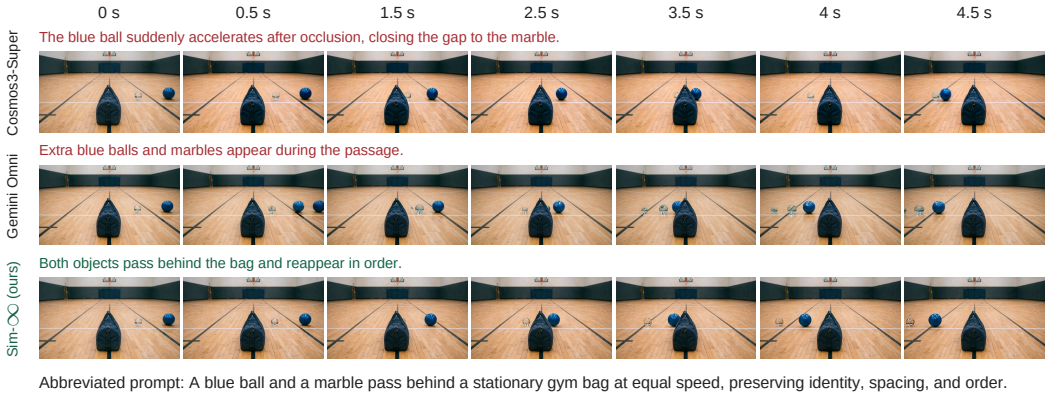


Figure 26: **Object permanence behind a gym bag.** A blue ball and a marble are prompted to pass behind a stationary bag at equal speed. Sim-$\infty$ preserves two separate objects through occlusion. Cosmos3-Super's blue ball suddenly accelerates after passing behind the bag, closing the gap to the marble; Gemini Omni introduces extra objects.

## C.2 COLLISION DYNAMICS

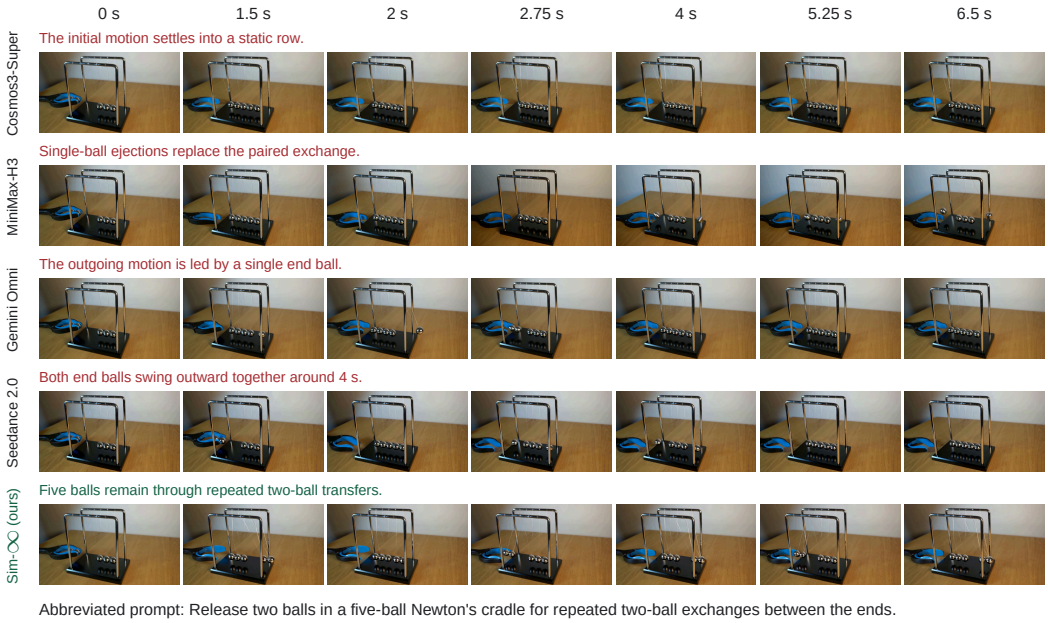


Figure 27: **Two-ball transfers in a five-ball Newton's cradle.** Cross-model comparison of two-ball exchanges in a five-ball Newton's cradle. The selected frames highlight outgoing-ball motion and object consistency.

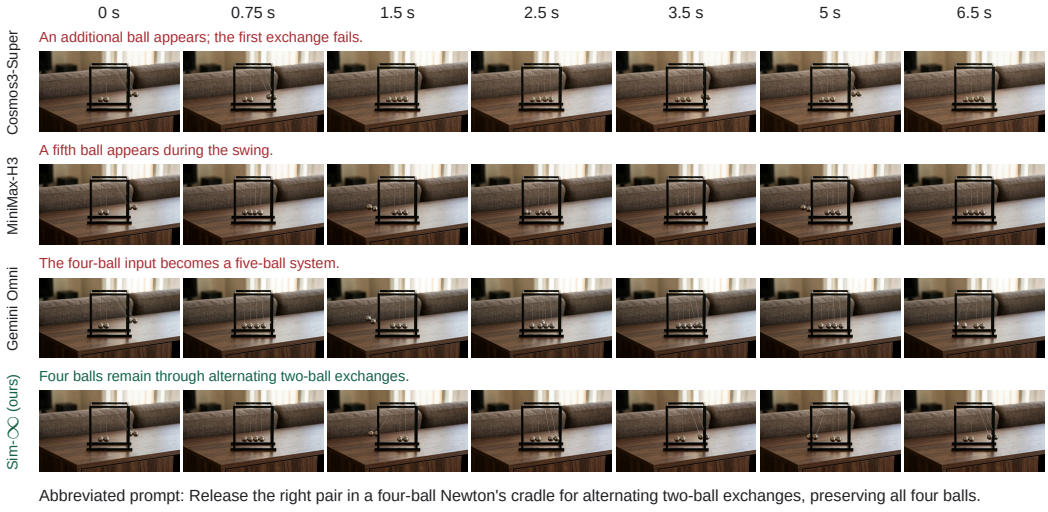


Figure 28: **Two-ball transfers in a four-ball Newton's cradle.** Sim-$\infty$ preserves the four-ball system through alternating two-ball exchanges, while all three compared models introduce an additional ball.

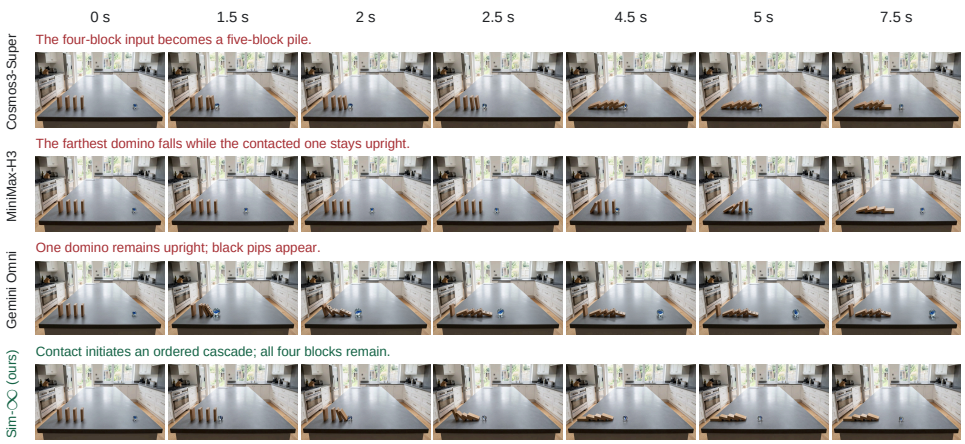


Figure 29: **Contact-triggered domino cascade.** A marble initiates a four-domino cascade. The comparison highlights contact-triggered toppling order and object consistency.

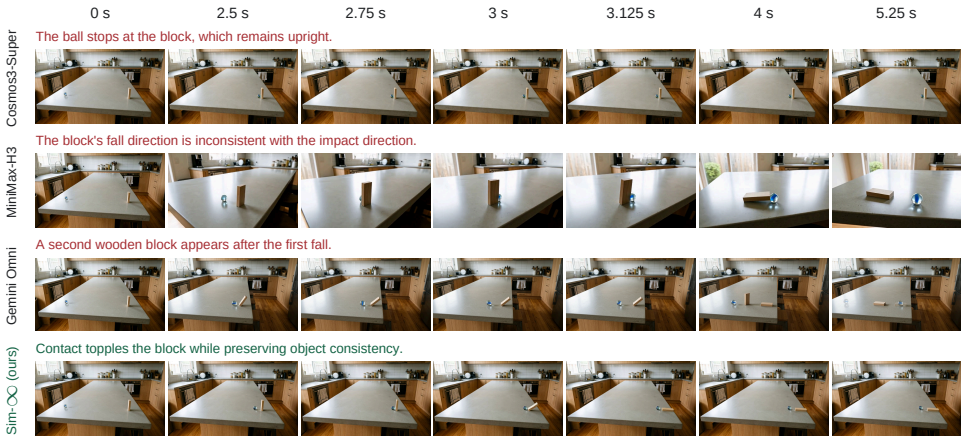


Figure 30: **Collision with a single wooden block.** A marble collides with one upright block. The comparison highlights the causal direction of toppling and object consistency.

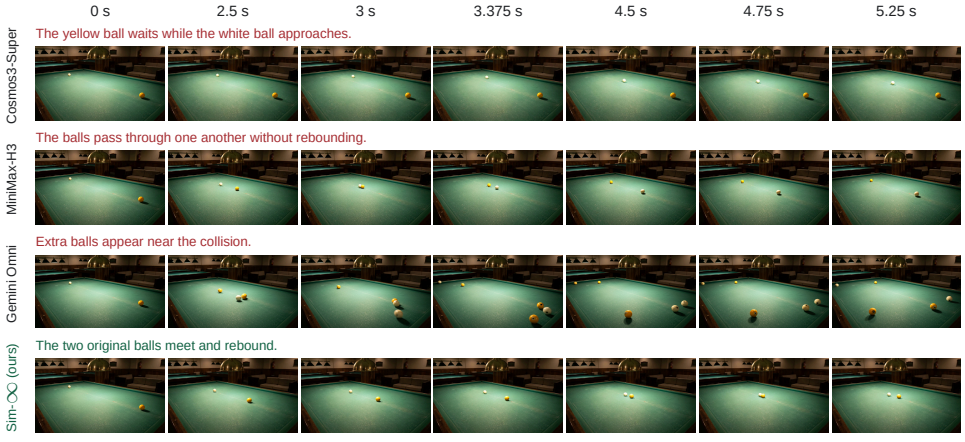


Figure 31: **Billiard-ball collision and rebound.** Two balls are prompted to approach and rebound after contact. Frames emphasize the encounter and post-collision response, using shared absolute times across models.

### C.3 WAVE PROPAGATION

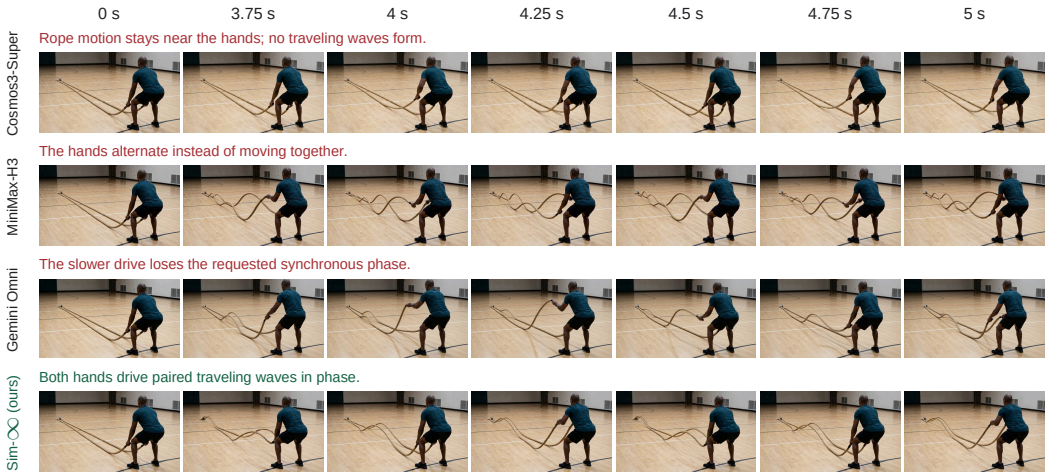


Figure 32: **Synchronous battle-rope motion.** Selected frames compare hand phase and wave propagation under a synchronous vertical drive.

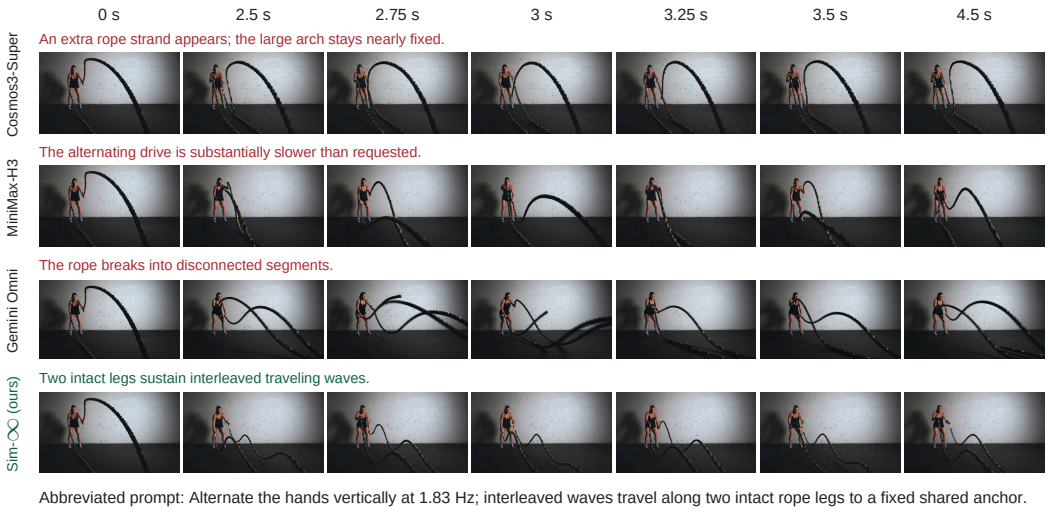


Figure 33: **Alternating battle-rope motion.** Selected frames compare interleaved traveling waves and rope continuity under an alternating vertical drive.

### C.4 CROSS-SCENE OVERVIEWS

Figures 34–36 show photorealistic generations from Sim-$\infty$ at a middle frame of the generated sequence.



Figure 34: **Newton's cradle across photorealistic scenes generated by Sim-$\infty$ at $t = 1.5$ s.** The supplementary video shows the complete generated sequences.

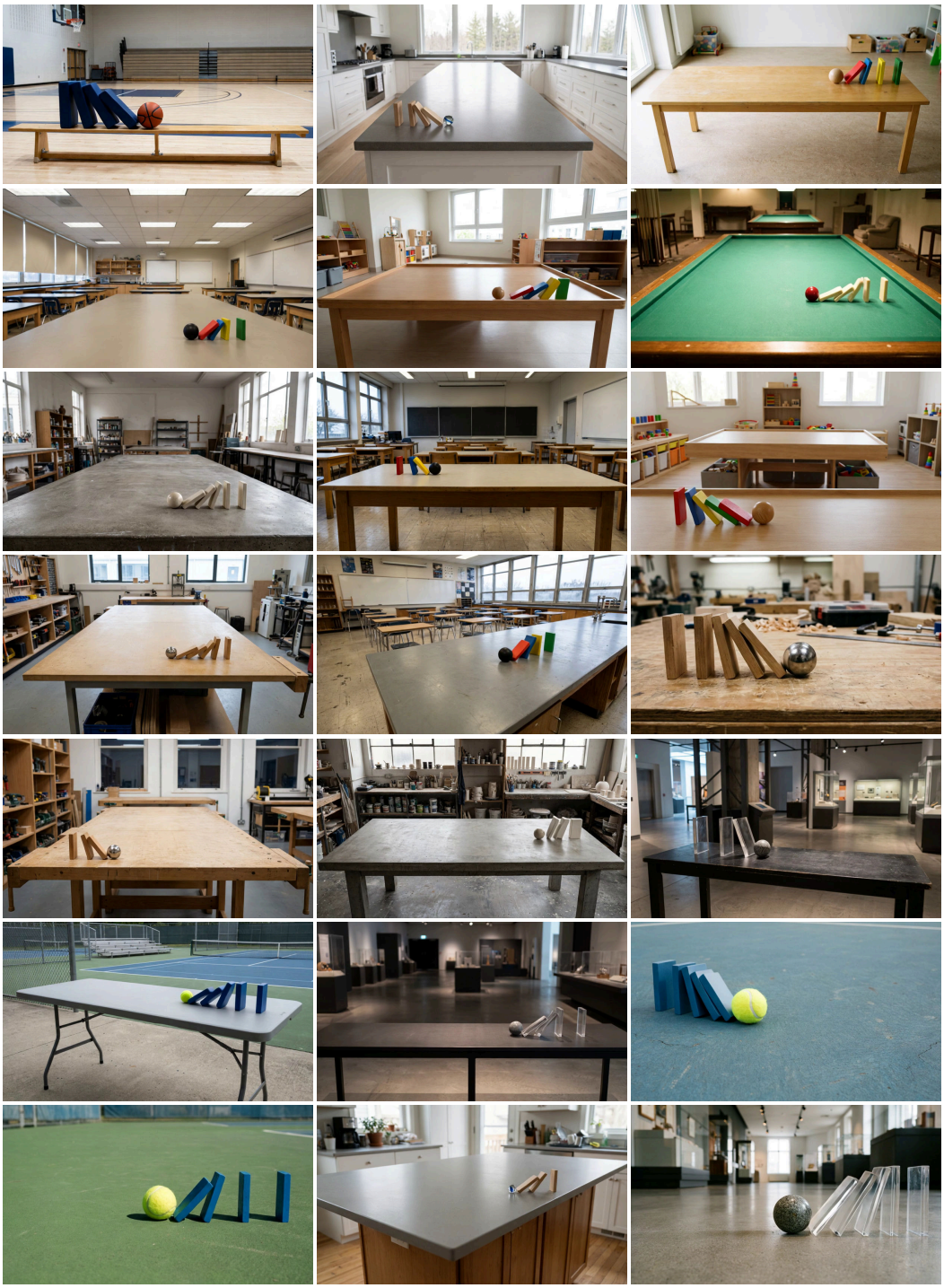

Figure 35: **Domino cascades across photorealistic scenes generated by Sim-$\infty$ sampled near each cascade.** Sampling times vary across panels to show contact and toppling configurations. The supplementary video presents the complete sequences at their original speed.

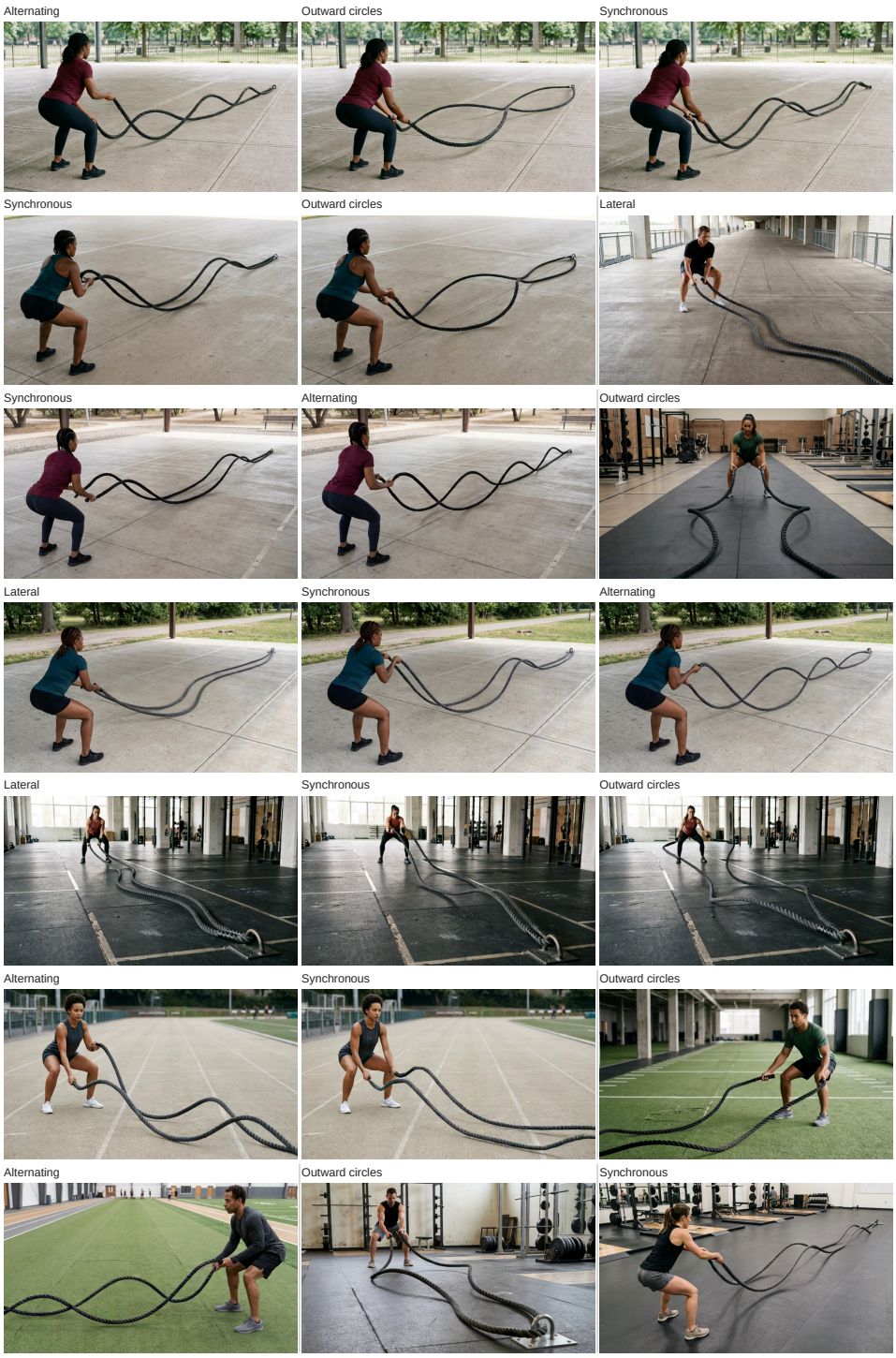


Figure 36: **Battle-rope motions across photorealistic scenes generated by Sim-$\infty$ at different timestamps.** Labels indicate the requested motion families, and examples sharing the same initial frame are placed together. The supplementary video shows their temporal evolution.

## D LIMITATIONS AND FUTURE WORK

**Evaluation metrics.** Evaluating physical fidelity remains challenging. At the judge's 8 FPS sampling rate, brief topology failures, such as momentary rope discontinuities, may go undetected. Although the judge agrees closely with human rankings when both give a strict ordering, small score differences should not be interpreted as precise differences in physical fidelity. Broader and more diverse physics benchmarks are also needed.

**Camera motion.** Our simulations use static cameras. Future work could investigate whether camera motion provides additional 3D cues that improve the learning of physical dynamics.

**Task coverage.** We study three physical skills. Extending this approach to other physical AI tasks, such as humanoid control, autonomous driving, and robotic manipulation, is a promising direction.