CRONOS: BENCHMARKING COUNTERFACTUAL PHYSICAL CONSISTENCY IN VIDEO MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video prediction is increasingly viewed as a path toward generalizable world models, yet it remains unclear whether these systems learn underlying causal structure or merely exploit superficial visual correlations for future prediction. We introduce CRONOS, an intervention-based benchmark to evaluate counterfactual physical consistency: whether a model's predictions of physical events respond appropriately to controlled changes in the visual input, such as variations of camera viewpoint, scene, object category, and object appearance. Built in a photorealistic Unreal Engine environment, CRONOS enables controlled, high-fidelity generation of videos across diverse scenes and dynamics, including sampling multiple physically plausible outcomes of the same physical event. In contrast to previous benchmarks, CRONOS systematically intervenes on four factors—camera viewpoint, scene, object category, and object appearance—while keeping the underlying physical event type, such as a fall, collision, or occlusion, fixed. Our evaluation of recent open-source video generators reveals substantial failures in counterfactual physical consistency: prediction quality for the same physical event type is affected by appearance, environment, and particularly viewpoint changes. CRONOS provides a controlled and reproducible testbed for diagnosing how generated video quality changes across interventions, establishing a concrete target for developing models that perform consistently across changes of multiple conditions. The dataset and evaluation code will be available upon acceptance.

1 INTRODUCTION

Recent progress in generative video modeling has made it increasingly plausible to learn *world models*—predictive models that capture how the visual world evolves over time and can support downstream reasoning and planning (Ha & Schmidhuber, 2018). Large-scale video diffusion models can synthesize temporally coherent, high-fidelity futures from partial observations, fueling the belief that scaling video prediction may yield generalizable predictive models of real-world dynamics (Ho et al., 2022b;a). However, visual realism alone does not imply that these predictive systems develop *causal representations* (Schölkopf et al., 2021) that capture relationships between objects, scenes, and dynamics, allowing robust predictions to remain stable under changes in viewpoint, appearance, or context. Such structured, causally meaningful representations are widely believed to be essential for robust generalization, compositional reasoning, and decision-making, as they enable models to distinguish underlying world dynamics from incidental visual correlations (Pearl, 2009; Richens & Everitt, 2024). However, it remains unclear whether modern models acquire such representations or primarily rely on superficial statistical regularities in the data for prediction. Studying this gap requires principled evaluations that move beyond perceptual quality and directly test whether a model's predicted future responds appropriately to controlled changes in the visual input.

Existing work has begun to probe whether video models capture physical and causal structure through specialized evaluation benchmarks. Some approaches construct controlled physics scenarios and assess predictions by comparing generated outcomes against ground-truth trajectories or physical constraints, measuring whether models obey expected dynamics such as collisions, motion, or conservation laws (Motamed et al., 2026; Zhang et al., 2025). Other methods rely on object-centric analyses, evaluating predicted trajectories or interactions using tracking and segmentation pipelines (Upadhyay et al., 2026; Li et al., 2025a), or employ vision–language models and human judgments to detect violations of physical plausibility (Bansal et al., 2025; Ma et al., 2026). While these

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.

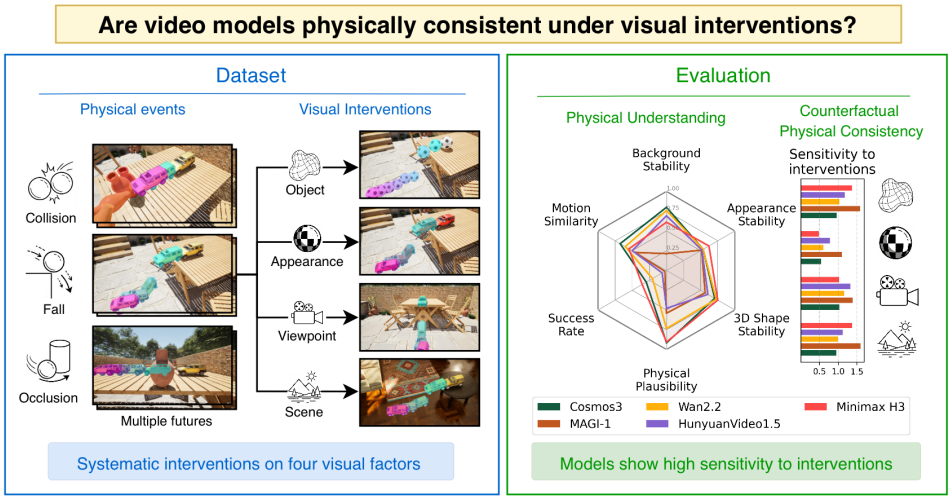

Figure 1: **The CRONOS benchmark** evaluates counterfactual physical consistency: whether a model's predictions of physical events respond appropriately to controlled changes in the visual input.

benchmarks provide insights into physical correctness, they largely evaluate predictions under a *fixed visual observation*. As a result, they reveal whether a model can produce a plausible continuation of a given scene, but provide limited insight into whether the underlying predictive representation is stable and structured. A reliable model should remain stable under nuisance changes such as viewpoint or appearance variations, while adapting coherently when other aspects of the scene change. We formalize this requirement through the notion of *counterfactual physical consistency*:

Counterfactual physical consistency refers to a model's ability to produce predictions of physical events that remain coherent across counterfactual variants of the visual input.

To study counterfactual physical consistency in modern video models, we introduce **CRONOS**, an intervention-based benchmark designed to evaluate how predictive video models respond to controlled changes in the visual world. CRONOS is built in a photorealistic Unreal Engine environment to enable the generation of realistic video sequences in which the underlying physical event type remains fixed while specific visual factors are systematically varied. In particular, we intervene along four complementary dimensions: camera viewpoint, scene, object category, and object appearance. Viewpoint and appearance changes primarily test robustness to nuisance variations that preserve physical parameters, while object-category and scene interventions probe whether models adapt coherently across changes in object properties and layouts. The benchmark spans three canonical interaction scenarios—rolling and falling, collisions, and occlusion and reappearance—chosen to isolate fundamental forms of basic physical interaction. The synthetic design allows CRONOS to also handle the challenge of physical parameter under-specification by sampling multiple physically plausible futures from the same conditioning signal. By explicitly controlling and recombining these factors, CRONOS enables analysis of counterfactual physical consistency in video models.

For evaluation, we introduce metrics that disentangle 3D motion from appearance, enabling a more fine-grained assessment of generation fidelity. Additionally, our intervention framework measures each model's sensitivity to controlled changes in the input signal, which serves as a diagnostic of counterfactual consistency. We apply these metrics to several state-of-the-art open-source video generation models under both image-to-video (I2V) and video-to-video (V2V) settings.

Our analyses reveal that models often fail to generate physically consistent videos and show substantial variation across intervention types, with especially high sensitivity for scene and viewpoint changes. Further, we show that video conditioning improves over image conditioning, and that scaling model size does not necessarily lead to more consistent generation quality. A preview of the dataset and the evaluation code can be found in the supplementary material. We will release both the code and the complete dataset upon acceptance. An overview of the data generation and evaluation in CRONOS can be found in Figure 1.

2 Related Work

Video generation models. Recent advances in video generation have produced models capable of synthesizing temporally coherent and visually detailed videos that are conditioned on text (T2V), images (I2V), past video frames (V2V), or combinations thereof. Early work extended image diffusion models to the temporal domain by inserting temporal layers into latent diffusion architectures (Blattmann et al., 2023b; Singer et al., 2022; Blattmann et al., 2023a; Ho et al., 2022a). More recently, transformer-based diffusion architectures (DiTs) (Ma et al., 2024) have enabled models such as CogVideoX (Yang et al., 2024), Wan (Wan et al., 2025), HunyuanVideo (Kong et al., 2024), and MovieGen (Polyak et al., 2024) to generate high-fidelity video at scale. Further, autoregressive formulations allow arbitrarily long generated sequences, as demonstrated by MAGI-1 (Teng et al., 2025) and Cosmos (Ali et al., 2025; Agarwal et al., 2026). However, despite advances in terms of visual fidelity, recent studies have shown that these models frequently violate basic physical principles such as object permanence, gravity, and cause-effect relations (Motamed et al., 2026; Kang et al., 2024; Zhang et al., 2025). This suggests that such models are limited in their ability to generalize physical understanding beyond visual patterns seen during training. While recent efforts (Li et al., 2025a) explored physics-aware post-training to mitigate such failures, these approaches still do not guarantee robustness. These findings highlight a crucial gap in current video generation models that CRONOS aims to evaluate systematically: counterfactual physical consistency, the capability of generating videos of physical events in consistent quality even when scene parameters change.

Evaluating video generation. Early evaluations of video generation models focused on image-based metrics to evaluate generation quality, such as FVD (Unterthiner et al., 2018), and were extended to capture various quality metrics (Huang et al., 2023; 2025; Liu et al., 2024; Feng et al., 2024). A growing set of benchmarks targets physical realism more directly where physical commonsense, physical laws, or scientific concepts are evaluated by humans, VLMs, or learned evaluators (Bansal et al., 2024; 2025; Meng et al., 2024; Chen et al., 2025; Gu et al., 2025; Guo et al., 2025; Hu et al., 2025; Li et al., 2025b; Zheng et al., 2025; Foss et al., 2025). Reference-based evaluations compare generations to trajectories, physical equations, real or simulated experiments (Li et al., 2025a; Motamed et al., 2026; Zhang et al., 2025; Upadhyay et al., 2026; Zhang et al., 2026). Specifically, PISA (Li et al., 2025a) compares object trajectories of videos that cover objects in free fall scenarios. Physics-IQ (Motamed et al., 2026) evaluates videos in real-world physical experiments through image-based metrics.

In contrast, Morpheus (Zhang et al., 2025) measures physics-informed scores of generated videos, specifically evaluating whether equations of motion and physical constraints are satisfied. WorldBench (Upadhyay et al., 2026) estimates physical parameters of generated videos based on simple real-world physical experiments and compares the results to synthetic videos acquired from a simulation environment. Although these reference-based evaluations move beyond perceptual metrics, each of them scores a generation against a single reference video. The initial frames of a scene typically under-specify its outcome, so a generated video may be physically valid yet differ from the one recorded continuation and be penalized for it. Following work on trajectory forecasting (Liang et al., 2020), CRONOS instead compares each generated video against a set of physically plausible reference videos, rewarding any outcome consistent with the underlying physics rather than one particular realization of it.

Across all discussed benchmarks, evaluated sequences are independent of each other and thus measure whether a single generation is physically plausible, rather than whether a model reacts correctly when the scene producing it changes, a perspective motivated by robustness evaluations (Hendrycks & Dietterich, 2019; Shu et al., 2020; Dünkel et al., 2025). CRONOS builds a dataset of sequences connected through controlled visual interventions, enabling the analysis of counterfactual physical consistency in video models by comparing the outcomes of different interventions.

Simulators for probing visual understanding. Synthetic environments enable controlled tests that are difficult to obtain from real videos. Many benchmarks make use of synthetic data in the realm of video reasoning: CRAFT (Ates et al., 2022), CLEVRER (Yi et al., 2020), and GRASP (Jassim et al., 2024) design pairs of questions and videos and evaluate models' understanding on simple scenes, while IntPhys (Bordes et al., 2025) focuses on detection of violations of physics. From the modeling perspective, Physion (Bear et al., 2021; Tung et al., 2023) evaluated different architectures' ability to predict the outcome of diverse physical events and PhysWorld (Kang et al., 2024) designed simple 2D environments to study generalization of visual properties on video diffusion. More recently, PISA (Li

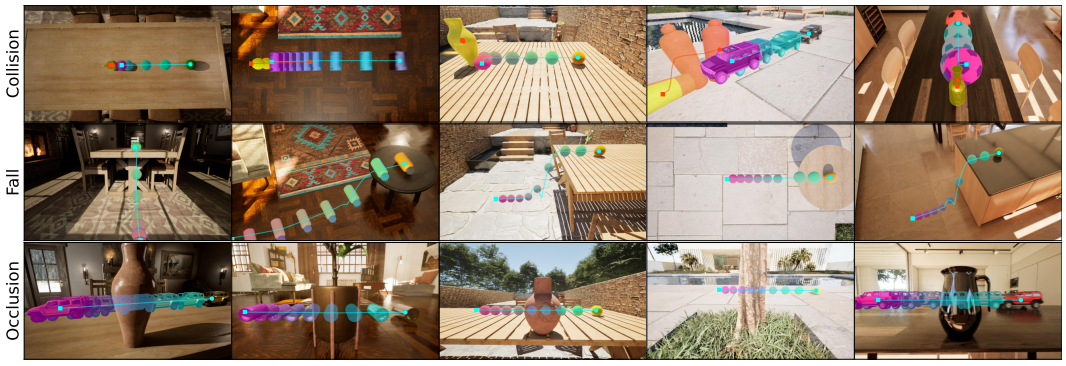

Figure 2: **CRONOS dataset overview.** Examples illustrating the three physical events (rows) in the CRONOS benchmark: collision, fall, and occlusion. For each event instance, we render multiple counterfactual observations by varying factors such as scene context, camera viewpoint, object category, and object appearance. Colored overlays show object trajectories across time, visualizing the underlying motion dynamics. This controlled design enables systematic evaluation of whether video models produce consistent predictions across different observations of the same event setup.

et al., 2025a) employed synthetic data to fine-tune and enhance physics modeling abilities of video models, while WorldBench (Upadhyay et al., 2026) generated synthetic scenes to evaluate physical understanding. Yet, most benchmarks leveraging synthetic data rely on objects with simple textures, and do not make use of high-fidelity rendering tools able to realistically simulate light. In contrast, CRONOS relies on a photorealistic simulator, keeping the advantages of a controlled environment while using higher-fidelity graphics than many synthetic physics benchmarks.

3 CRONOS BENCHMARK

CRONOS frames the evaluation of video generation models as a controlled counterfactual experiment. The core experimental unit in CRONOS is a *physical event*: a basic physical simulation specified via initial states, impulses, and simulator parameters that defines a scene's 3D dynamics. From each event type, we render a set of *counterfactual observations* by intervening on a single factor at a time—camera viewpoint, scene, object category, or object appearance—while holding the remaining variables fixed. Some interventions preserve the underlying physical parameters, such as viewpoint and appearance, while others change contextual or object-level properties that may alter the expected rollout. This design enables measurement of *counterfactual physical consistency*: whether model predictions remain stable under nuisance interventions that do not alter the event dynamics (e.g., viewpoint) and vary coherently when interventions induce structured changes (e.g., object class). The remainder of this section describes our controlled simulation pipeline for generating event instances (Section 3.1), the canonical physical events defining underlying dynamics (Section 3.2), the systematic intervention protocol used to render counterfactual observations (Section 3.3), and the object-centric metrics quantifying prediction accuracy and intervention sensitivity (Section 3.4).

3.1 DATA GENERATION

We generate all sequences in a controllable Unreal Engine environment (Epic Games, 2025). Each event is specified by carefully selected simulator configurations, allowing targeted interventions of individual factors of realistic events. This control is difficult to obtain from real video, where camera viewpoint, object appearance, scene context, and dynamics cannot be independently varied while preserving the same physical event type. To consider a range of physically plausible outcomes, we generate three variants of each sequence by selecting different physical parameters. All scenes are rendered at 1920×1080 pixels and 30 FPS using high-quality professional 3D assets chosen to reflect common real-world environments, including indoor and outdoor environments under diverse lighting conditions. In addition to RGB frames, the simulator provides per-object segmentation masks, used for the object-centric metrics in Section 3.4. We show examples of rendered scenes for all physical events in Figure 2. Additional details about the dataset can be found in Section A.

3.2 PHYSICAL EVENTS

CRONOS uses three physical events that probe complementary aspects of predictive reasoning while keeping the setup compact. All are generated from standardized initial conditions in which an impulse initiates object motion. Differences across intervention variants therefore come from the controlled visual change, not a new setup. We consider three scenarios:

Fall. An object rolls across a surface and falls from an edge, testing prediction across changing contact conditions and free-fall motion.

Collision. One object impacts another, testing whether generated videos preserve physically plausible interaction dynamics, including temporal and spatial coherence and object permanence.

Occlusion. An object rolling across a smooth surface becomes fully occluded behind another object and later reappears, testing whether models capture long-range temporal coherence.

Together, these events provide controlled yet diverse dynamic settings for the systematic analysis of counterfactual consistency via the interventions introduced next.

3.3 SYSTEMATIC VISUAL INTERVENTIONS

Building on the controlled simulation setup (Section 3.1) and physical event dynamics (Section 3.2), CRONOS systematically renders a set of interventions. For sensitivity analysis, we group variants that differ along one intervention axis while holding the remaining variables fixed:

Scene intervention. The background environment and scene layout are changed (e.g., height in fall sequences), testing whether models remain reliable across contextual changes and adapt to layout-dependent dynamics when scene geometry affects the rollout.

Camera viewpoint intervention. The rendering viewpoint is changed while keeping scene dynamics intact, probing whether models disentangle scene geometry from observed motion while maintaining perspective consistency.

Object appearance intervention. Visual object attributes, such as color, are changed without altering physical parameters, isolating whether models disentangle appearance from dynamics.

Object-category intervention. The target object is replaced with a compatible one, changing both visual properties (e.g., shape, material) and physical parameters (e.g., mass, friction), which directly affect the dynamics. This intervention probes whether models adjust predictions coherently across object instances whose visual and physical properties differ, or instead rely on object-specific correlations learned during training.

The dataset follows a full-factorial design for each physical event, except viewpoint, which is fixed for occlusion events to preserve the intended visibility structure. This enables fine-grained analysis of sensitivities and counterfactual consistency in generated videos.

3.4 EVALUATION METRICS

CRONOS decomposes generation quality into complementary per-video metrics—background stability, motion similarity, appearance stability, 3D shape stability, and physical plausibility—together with a global success criterion that aggregates them into a single pass/fail signal per video. Physical parameter estimation from one or a few frames is inherently ill-posed, so it would be unfair to penalize models by assuming a single correct rollout. During evaluation, CRONOS therefore compares generated videos to all available rollouts, and selects the best match. All metrics are normalized to $[0, 1]$ and higher values always indicate superior performance. Additional details regarding metric computation and other steps, including segmentation masks, visibility filtering, aggregation rules, and thresholds are described in Section C.

Background stability measures whether background regions remain coherent and fixed relative to the conditioning frame by computing pixel-wise error, following WorldBench (Upadhyay et al., 2026). It captures artifacts such as background morphing, lighting drift, camera motion, and new objects, all undesired and explicitly excluded by the text prompt (Table 4).

Motion similarity compares the motion embeddings of multiple reference and the generated frames using the appearance-invariant motion encoder from DisMo (Ressler-Antal et al., 2025).

Appearance stability measures whether each object preserves its visual identity over time, using cosine similarities of per-object DINOv2 (Oquab et al., 2023) embeddings against the initial frame, as in VBench (Huang et al., 2023). CLS tokens are computed from object-masked images.

3D shape stability measures whether object geometry changes by comparing per-frame SAM3D-reconstructed object meshes to the initial mesh via the Chamfer distance.

Physical plausibility measures high-level event correctness and physical violations using a VLM-as-judge protocol (Ma et al., 2026; Zheng et al., 2025) with Qwen3-VL-32B (Bai et al., 2025). This fixed set of binary questions covers common physical violations and event-completion criteria.

Success rate aggregates per-video metrics into a binary pass/fail criterion. A video is successful only if all quality metrics pass their calibrated thresholds and no object disappearance is detected. Thresholds are calibrated from the human study in Section 4.2 by requiring equal ratios of false positive and false negative rates, where failed videos received low annotator ratings for the corresponding metric. Further, the disappearance detector prevents segmentation failures from producing artificially high object-centric scores. The success rate is the fraction of videos passing this test.

Sensitivity to interventions. We measure how much each intervention changes the quality of generated videos. For groups of videos differing only along one intervention axis, we compute the average pairwise difference of each metric, normalize it by the average pairwise difference of that metric across seeds (shared across all models), and average across groups, seeds, and metrics. A lower sensitivity indicates higher **counterfactual physical consistency**. Note that the sensitivity measures consistency in quality rather than similarity between outputs themselves.

4 RESULTS

In the following, we present the model evaluations using our framework and discuss our findings. After clarifying the experimental setup in Section 4.1, we present the user study used to validate the selected metrics in Section 4.2 before discussing the benchmark's findings in Section 4.3. Finally, we study the simulation-to-real gap with a real data experiment in Section 4.4.

4.1 EXPERIMENTAL SETUP

Our evaluation includes several state-of-the-art open-source video generation models: Cosmos2.5 (Ali et al., 2025), Cosmos3-Super (Agarwal et al., 2026), CogVideoX1.5 (Yang et al., 2024), MAGI-1 (Teng et al., 2025), Wan2.2 (Wan et al., 2025), HunyuanVideo1.5 (Wu et al., 2025), and MiniMax-H3 (MiniMax AI, 2026). For Cosmos2.5, we evaluate the two available model sizes 2B and 14B to study the effect of model scaling in video generation models. We run all models in the I2V setting using a single frame, and Cosmos and MAGI-1 in the V2V setting, using a five-frame clip, which reveals the initial direction and velocity while still requiring future prediction capabilities.

We use each model's standard settings and the same event-specific prompt template that describes the scene configuration, contextual details, and intended motion. Detailed text prompts can be found in Section B and qualitative examples of generated videos are shown in Section E. In order to consider model variance, we sample three seeds per experiment and average the metrics across all seeds. For sensitivities, we compute the sensitivities across intervened experiments with the same seed, and then average across seeds, groups, and metrics.

4.2 HUMAN EVALUATION STUDY

To validate our selected metrics, we perform a user study that rates generated videos along the evaluated dimensions. Specifically, we hire qualified annotators via Prolific and ask them to label the quality of *object appearance*, *object shape*, *background stability*, *motion plausibility*, and *event quality*. We follow Bansal et al. (2025) and evaluate on a scale between 1 (very poor) and 5 (excellent). We select 540 representative videos from various models with diverse objects and scenes, and we collect median-aggregated ratings of three annotators. Note that not all evaluated models are included in the study, as the most recent models where released after the study was conducted. More details including

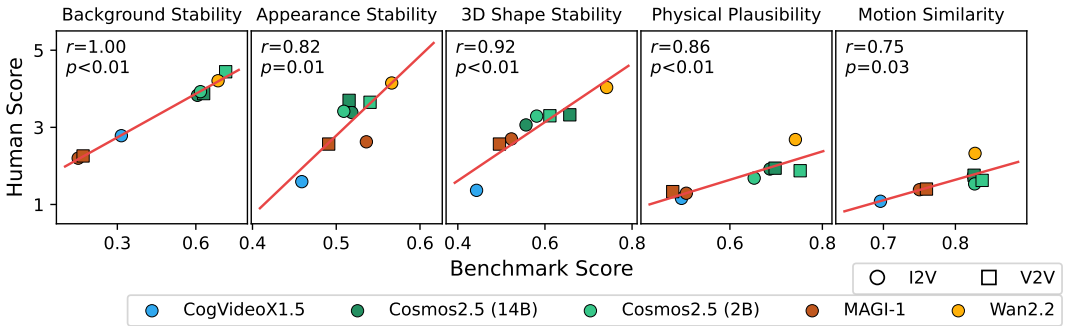

Figure 3: **Results of the human evaluation study.** Model performances positively correlate with human ratings (higher is better), as the Pearson correlation coefficients indicate.

annotation instructions are provided in Section D. We provide Pearson correlation coefficients for measured performance vs. human rating in Figure 3. The stable and positive correlations support using the proposed metrics for the subsequent analysis on the complete benchmark. We also verify the stability of the success-rate thresholds in Section C.

4.3 Analysis of Benchmarking Results for Video Models

Physical event generation. The averaged metrics in Table 1 show that performance differs substantially across models, with Cosmos3-Super and MiniMax-H3 performing best and MAGI-1 and CogVideoX1.5 being clearly worse. Since a video is only successful if all metrics pass their calibrated thresholds, success rates are low: Cosmos3-Super (V2V) and MiniMax-H3 reach 0.30 and 0.36, while all other models range from 0.01 to 0.23. Human ratings agree: average event quality stays below 2 on the 1–5 scale for all annotated models except Wan2.2 (2.68, Table 10). Failure profiles also differ across models (Figure 12): Cosmos3-Super rarely fails physical plausibility but often fails appearance stability (45–55% of fall and collision videos), MiniMax-H3 has the lowest appearance-stability failure rates but frequently fails background stability, and HunyuanVideo-1.5 keeps backgrounds stable but fails motion similarity in up to 88% of videos. No single quality dimension therefore explains the low success rates.

Finding #1: All evaluated open-source video models struggle to reliably generate short clips of basic rigid-body physics. The strongest evaluated model achieves only 36% success rate, with most models below 20%.

Visual stability versus physical correctness. High visual quality does not imply physical correctness. On the fall event (Table 8), Wan2.2 achieves the best background stability and motion similarity of all models, yet its physical plausibility (0.47) falls below the calibrated threshold, which only Cosmos3-Super and MiniMax-H3 exceed. Similarly, HunyuanVideo-1.5 produces comparatively stable backgrounds but has the lowest physical plausibility overall (0.48). These models generate clean and stable videos of events that do not unfold as physically expected.

Finding #2: Visual stability does not imply physical correctness: models that generate stable backgrounds and objects can still fail to produce the expected physical event, most clearly for falling objects.

Counterfactual consistency. Figure 4 reports each model's sensitivity to interventions. A sensitivity below 1 indicates that an intervention changes quality less than resampling the seed, and above 1 that it changes quality more. Since the normalization is shared across models, sensitivities are directly comparable between them. Consistency is closely tied to benchmark performance: across the 12 model–mode configurations, success rate and average sensitivity are strongly anti-correlated (Spearman $\rho = -0.87$), with Cosmos3-Super and Wan2.2 being the most consistent and MAGI-1 and CogVideoX1.5 the most sensitive. MiniMax-H3 is an exception: it is the most consistent model under object and appearance interventions but highly sensitive to viewpoint and scene changes, indicating a highly pixel-based world encoding.

Across models, appearance changes are tolerated best, although for MAGI-1 and CogVideoX1.5 even recoloring the object changes quality at least as much as resampling the seed (Table 9). Scene and

Table 1: **Benchmark performance averaged across all videos.** Metrics are normalized to $[0, 1]$ and higher is better. Best, second-, and third-best values per metric are indicated by bold, underline, and dashed underline, respectively. A video is counted as successful only if every per-video metric passes its calibrated threshold. The general low success rates indicate that meeting all quality criteria simultaneously remains challenging for current open-source video models.

Mode	Model	Bg Stab.	Motion Sim.	App. Stab.	3D Shape Stab.	Physical Plaus.	Success Rate
V2V	Cosmos2.5-2B	0.77	0.64	0.53	0.67	0.71	0.23
	Cosmos2.5-14B	0.55	0.56	0.51	0.64	0.67	0.14
	Cosmos3-Super	**0.80**	**0.68**	0.54	0.71	**0.92**	0.30
	Magi-1-4.5B	0.22	0.47	0.47	0.58	0.51	0.01
I2V	Cosmos2.5-2B	0.60	0.56	0.50	0.63	0.66	0.12
	Cosmos2.5-14B	0.52	0.52	0.49	0.60	0.67	0.08
	Cosmos3-Super	0.71	0.63	0.50	0.70	0.91	0.19
	Magi-1-4.5B	0.21	0.46	0.51	0.57	0.54	0.02
	CogVideoX1.5-5B	0.39	0.35	0.40	0.49	0.58	0.02
	Wan2.2-14B	0.76	0.60	0.54	0.73	0.73	0.20
	HunyuanVideo-1.5	0.69	0.50	0.52	0.61	0.48	0.06
	MiniMax H3	0.61	0.55	**0.62**	**0.75**	0.91	**0.36**

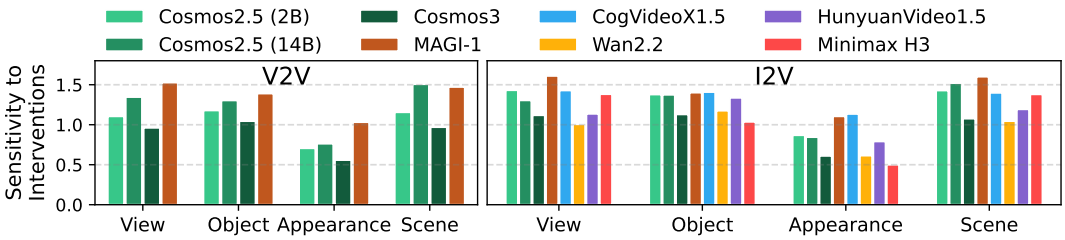

Figure 4: **Sensitivity to counterfactual interventions.** Sensitivities are averaged across metrics, and lower values indicate lower sensitivity. All evaluated models show substantial variation across intervention types, including appearance changes that alter only objects' visual properties.

object interventions induce larger variation. While they change the expected rollout, a consistent model should still generate each variant with similar quality, as each video is evaluated against its own references. Viewpoint changes preserve the dynamics, object properties, and scene, yet they also induce large variation in all models. This indicates that current open-source models encode predictions in a view-dependent way, relying on visual statistics that do not transfer across viewpoints.

Finding #3: Models do not achieve robust counterfactual physical consistency: generation quality changes substantially across interventions, especially for scene and viewpoint changes, and the models with the highest success rates are generally the least sensitive.

Effect of video conditioning. Comparing I2V and V2V for Cosmos and MAGI-1 (Table 1), V2V outperforms I2V in most metrics. Beyond the expected gain in motion similarity, as the conditioning frames reveal the direction and magnitude of motion, video conditioning also improves background stability for all these models and object stability in both appearance and shape for the Cosmos family. We hypothesize that conditioning frames help the model build more robust object representations, and that the static camera in the input clip is a stronger motion cue than text alone. Video conditioning also improves consistency: average sensitivity is lower in the V2V setting for 10 of the 12 model–event pairs and unchanged for the other 2 (Table 9), as the observed motion, which is shared by all intervention variants, anchors the prediction independently of the intervened visual factors.

Finding #4: Video conditioning improves not only motion fidelity, as expected from the additional temporal signal, but also the stability of backgrounds and objects and the consistency of generation quality across interventions.

Effect of model size. Comparing Cosmos2.5-2B and Cosmos2.5-14B isolates the effect of parameter count within one model family. Surprisingly, the 14B variant performs worse on nearly every metric in both settings (Table 1), with the V2V success rate dropping from 23% to 14%. The

largest gap is in background stability, while physical plausibility is comparable and human annotators rate the motion of the 14B variant slightly higher (Table 10). Consistency does not improve either: the 14B variant is more sensitive than the 2B variant on all three events in V2V and on par in I2V (Table 9). This agrees with Kang et al. (2024), who show in a simplified 2D setting that larger video generators do not necessarily learn physical laws that generalize out of distribution, and extends this observation to photorealistic scenes and modern pretrained models. Cosmos3-Super outperforms both Cosmos2.5 variants, but this gain cannot be attributed to scale alone, as architecture and training data also change. Since our comparison covers a single model family and scale step, it motivates a more detailed scaling study, which our benchmark enables.

Finding #5: Scaling Cosmos2.5 from 2B to 14B parameters improves neither generation quality nor consistency under interventions, indicating that model size alone does not guarantee counterfactual physical consistency.

4.4 Real data experiment

CRONOS makes use of photorealistic renders in order to precisely control visual factors necessary for the interventions setup. This opens the question of the simulation-to-real gap, i.e, whether findings from our simulated setup transfer to real world data. We study the effect in real world in a smale scale experiment. In particular, we record a number of sequences imitating the physical events in CRONOS's main dataset. Quantitative results suggest that the gap is small, as metrics stay in the same range, consistent with the findings of (Upadhyay et al., 2026). Qualitatively, we observe similar failure cases. We present metrics and individual rollouts in Section F.

5 Limitations and Future Work

Synthetic-to-real domain gap. CRONOS uses Unreal Engine renderings rather than real videos. This control is necessary for matched counterfactual interventions and is common in synthetic physics benchmarks (Yi et al., 2020; Bordes et al., 2025; Bear et al., 2021), but it introduces a domain gap despite the high visual fidelity of the selected scenes. Experiments with real data (Section F) suggest that the failure modes and general trends are shared across data domains.

Benchmark scope. CRONOS covers three basic rigid-body interactions, chosen as canonical primitives that isolate single- and multi-object dynamics. Even the strongest evaluated model succeeds on only 36% of videos. So, harder phenomena would add difficulty without making failures easier to diagnose. Natural extensions include further rigid-body events (ramps, stacking, containment) that reuse the current metrics unchanged, and a non-rigid track with its own evaluation.

Scope of evaluated models. We evaluate open-source models with fixed weights and reproducible settings, not closed commercial systems such as Veo, Sora, or Kling. We purposefully avoid evaluating closed models to ensure scientific reproducibility and verifiability by the research community. On top of that, select open-source models often outperform proprietary models in common physics benchmarks (Motamed et al., 2026; Rädsch et al., 2026).

6 Conclusions

We introduced CRONOS, an intervention-based benchmark for evaluating counterfactual physical consistency in video generation models. CRONOS consists of high-quality synthetic videos of controlled physical events, including falling dynamics, collisions, and occlusions. For each event, we systematically intervene on four visual factors—camera viewpoint, scene, object category, and object appearance—enabling fine-grained analysis of model sensitivity to structured changes. To evaluate predictions, we combine state-of-the-art foundation models, handcrafted rules, and a VLM-as-judge protocol to design a collection of metrics focusing on object stability and physical plausibility that are validated through a user study. Our experiments reveal substantial limitations of evaluated open-source models: even the strongest among them struggle to generate simple physical events and show substantial inconsistencies under targeted visual interventions, revealing low counterfactual physical consistency. Our findings suggest that many models rely on superficial visual correlations rather than producing stable predictions of scene dynamics. In general, CRONOS provides a useful testbed for diagnosing these limitations and guiding the development of video models that produce more robust and structured predictions of the visual world.

AI DISCLOSURE

In this work, we have used generative AI tools (coding agents) for implementing methods, in particular parts of the code on the evaluation. We have manually reviewed and tested all generated code. We have not used generative AI tools for generating synthetic datasets, developing frameworks, formulating mathematical claims, writing proofs, designing experiments, assisting with translation, formatting datasets, supporting analyses, or interpreting results. We have also used generative AI tools for formulating parts of the user survey, and drafting sections of the paper. These fragments have only been used as a starting point and manually refined. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Human study. We used external annotators recruited through Prolific to gather human scores for the evaluation. Annotators participated voluntarily and were informed about the nature of the task before participation. They were compensated at a fair rate for their time and effort, in accordance with Prolific's payment guidelines. Additional information regarding the study can be found in Section D.

Asset license. 3D assets were purchased to create the dataset. We do not distribute the assets themselves, but we provide the rendered video frames and metadata. Additional details regarding the licenses and documentation can be found in Section H.

REPRODUCIBILITY STATEMENT

We provide the evaluation code as well as a preview of the dataset in the supplementary material. The full dataset will be released upon acceptance. We provide additional details regarding the dataset design in Section A and evaluation in Section C.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

Tayfun Ates, M. Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. CRAFT: A benchmark for causal reasoning about forces and inTeractions, 2022. URL https://arxiv.org/abs/2012.04293.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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. *arXiv preprint arXiv:2406.03520*, 2024.

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

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Yongfan Chen, Xiuwen Zhu, and Tianyu Li. A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction, 2025. URL https://arxiv.org/abs/2502.05503.

Olaf Dünkel, Artur Jesslen, Jiahao Xie, Christian Theobalt, Christian Rupprecht, and Adam Kortylewski. CNS-Bench: Benchmarking image classifier robustness under continuous nuisance shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19978–19988, 2025.

Epic Games. *Unreal Engine*, 2025. URL https://unrealengine.com. Software.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024.

Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. CausalVQA: A physically grounded causal reasoning benchmark for video models, 2025. URL https://arxiv.org/abs/2506.09943.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A comprehensive evaluation of physical realism in text-to-video models, 2025. URL https://arxiv.org/abs/2507.13428.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2VPhysBench: A first-principles benchmark for physical consistency in text-to-video generation, 2025. URL https://arxiv.org/abs/2505.00337.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Jonathan Ho, William Chan, Chitwan Saharia, Mohammad Norouzi, and David Fleet. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022b.

Lanxiang Hu, Abhilash Shankarampeta, Yixin Huang, Zilin Dai, Haoyang Yu, Yujie Zhao, Haoqiang Kang, Daniel Zhao, Tajana Rosing, and Hao Zhang. Benchmarking scientific understanding and reasoning for video generation using VideoScience-Bench, 2025. URL https://arxiv.org/abs/2512.02942.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Ziqi Huang et al. VBench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. GRASP: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6297–6305, 2024.

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. *arXiv preprint arXiv:2411.02385*, 2024.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6013–6022, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Yao, Guanyu Zhu, Tony Fang, Hao Wu, Yatian Ai, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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 *International Conference on Machine Learning*, pp. 35685–35709. PMLR, 2025a.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. WorldModelBench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025b.

Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10505–10515. IEEE, 2020.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? evaluating state evolution in video world models. *arXiv preprint arXiv:2603.13215*, 2026.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

MiniMax AI. MiniMax H3: An open-weight, general-purpose, omni-modal generation model, 2026. URL `https://www.minimax.io/blog/minimax-h3`.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Anurag Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

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.

Thomas Ressler-Antal, Frank Fundel, Malek Ben Alaya, Stefan Andreas Baumann, Felix Krause, Ming Gui, and Björn Ommer. Dismo: dismotangled motion representations for open-world motion transfer. *arXiv preprint arXiv:2511.23428*, 2025.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Michelle Shu, Chenxi Liu, Weichao Qiu, and Alan L. Yuille. Identifying model weakness with adversarial examiner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11998–12006. AAAI Press, 2020. doi: 10.1609/aaai.v34i07.6876.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Pranay Boreddy, Shruti Satya Narayana, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. WorldBench: Disambiguating physics for diagnostic evaluation of world models. *arXiv preprint arXiv:2601.21282*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

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

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: CoLlision events for video REpresentation and reasoning. In *International Conference on Learning Representations*, 2020.

Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments, 2025. URL https://arxiv.org/abs/2504.02918.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-Eval: Evaluating physical realism in generated video via human reasoning, 2026. URL https://arxiv.org/abs/2603.19607.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Supplementary Material

A Dataset Details

The dataset of CRONOS is handcrafted using combinations of object and scene assets. Objects are selected to align with the physical events under study, ensuring that their geometry and material properties allow sliding or rolling over smooth surfaces, thereby producing visually plausible and dynamically rich motion patterns.

Simulated physical parameters, including mass, friction, and restitution coefficients, are explicitly defined per object and are kept consistent across events. In order to render multiple variations of each possible configuration, we select key physical parameters of the physical events and modify them. In particular, we select a high, medium and low value for: moving object mass (collision), elasticity (fall) and friction (occlusion). This way, we cover the range of plausible physical behaviors, while preserving the systematic interventions intact.

Scene layouts and initial conditions (e.g., object positions and applied forces), are individually tailored to fit each physical event, and camera viewpoints are strategically positioned to preserve global context while maintaining clear visibility of the relevant interactions throughout the sequence. This ensures that differences between rendered videos are attributable only to controlled interventions on the scene state. The proposed controlled generation process forms the basis for the construction of matched counterfactual observations used throughout CRONOS. Table 2 shows the number of different variations per intervention type. An overview of the original asset scenes, objects used for interventions and their different appearances are shown in Figure 5. For all three physical events, all variations are rendered. The only exception is the occlusion event, where the view is not altered to preserve the intended visibility effect. The number of videos rendered per event is shown in Table 3.

Table 2: **Number of variations per intervention.**

Intervention	# of variations
Scene	5
Object	5
Appearance	3
View	4

Table 3: **Number of videos per physical event.**

Event	# of videos
Fall	900
Collision	900
Occlusion	225
Total	2025

B Text Prompts

In addition to the visual conditioning signal, text prompts are provided to all video models. These prompts have been manually designed for each physical event, and follow the same structure across interventions, as shown in Table 4. The prompts coarsely describe each physical event, serving as an overview of the expected dynamics but without providing many details. The specific variation values of each intervention type are added through the prompt, as well as details required by each physical event (e.g., the collision target and the occluding object). On top of this, an additional shared prompt is added in all cases, providing general details about the expected physical behavior and camera movement.

C Evaluation Details

Robust minimum and maximum estimation. Observing generated videos, we realized that alterations or hallucinations in a few frames can completely change the way humans perceive the video. Therefore, aggregating per-frame results using the average is not appropriate, as the effect of a brief but meaningful error gets diminished. On the other hand, considering only the worst frame is unstable and can be affected by limitations of our metrics. Instead, we compute the average of the worst k frames in the video, an aggregation we denote RobustMin (or RobustMax, for metrics



Figure 5: **Dataset assets.** Overview of Unreal Engine assets used for all rendered sequences: original scenes, object models, and appearance variations.

Table 4: **Structure of text prompts.** General text prompt structure for each physical event. Variables in brackets are substituted using the details of each particular video. An additional prompt describing the general physical behavior and expected movement is added to all physical events.

Event	Prompt Structure
Fall	The video shows a {view} view of a {appearance}{object} smoothly {object_movement} across a {surface} in a {scene}. When the {object} reaches the edge, it falls off the {surface} and vertically descends until it hits the ground. The object maintains its shape and does not break during the fall.
Collision	The video shows a {view} view of a {appearance}{object} smoothly {object_movement} across a {surface}, colliding with a {collision_target} in a {scene}. The objects react to the collision but preserve their rigidity and do not break.
Occlusion	The video shows a {view} view of a {appearance}{object} smoothly {object_movement} and passing behind a {occluding_object} on a {surface} in a {scene}. The {object} becomes temporarily occluded by the {occluding_object} before reappearing. The {occluding_object} remains stationary throughout the sequence.
Additional prompt	Everything in the video follows the natural behavior of solid objects in a physical environment. Objects do not fly, morph, or disappear. No new elements appear in the scene. The background is static. Fixed camera view, no camera movement.

where larger values are worse) in the definitions below. We select k to be around 5% of the total generated frames. In cases where we compute scores per object, we select the worst k scores across both sets of values. This way, videos where a single object behaves unrealistically do not get their scores averaged out by the other object.

Object segmentation masks and point tracks. For all videos, we estimate per-object segmentation masks using SAM3 (Carion et al., 2025) and point tracks using CoTracker3 (Karaev et al., 2025). In both cases, we prompt the model with the original rendered masks in the first frame, which always matches the ground truth renders. We then propagate through the video to obtain per-frame predictions.

Object disappearance detection. Videos in which objects vanish abruptly can yield artificially inflated scores on object-centric metrics, since evaluation is only meaningful on frames where the target object is present. At the same time, given our video settings, we must account for legitimate effects such as occlusion and objects leaving the scene. To disentangle these cases, we introduce an object disappearance detector. A disappearance is registered only when two criteria are jointly satisfied: (i) the object is absent from the disappearance frame onward, and (ii) the object does not exit the image boundary at the disappearance time. We assess the first criterion using the predicted segmentation masks and the second using object tracks, which extrapolate motion beyond the visible frame. Because very rapidly vanishing objects might artificially obtain high scores in some metrics, we conservatively assign videos containing disappearing objects the minimum score on appearance stability, 3D shape stability, and motion similarity.

Occlusion filter. Beyond disappearance detection, object-level metrics (appearance stability and 3D shape stability) operate on segmentation masks and therefore require a sufficient number of visible pixels to be reliable. Under severe occlusion, these metrics become unstable. To mitigate this, we introduce an occlusion filter that restricts metric computation to frames with low occlusion. We

estimate the occlusion level of an object as the ratio between its current mask size and its mask size in the first frame. We find this simple heuristic to be effective in practice, as the apparent size of objects remains approximately constant across most videos in our setting. For each object, we retain only frames in which the estimated visible fraction exceeds a threshold of 25% of the initial mask size, and we apply this selection independently per object.

Background stability. We quantify alterations on the background, which should remain static in all sequences. This way, we are able to detect morphing, new objects appearing, and camera movement. We measure changes by comparing all generated frames to the initial reference frame via pixel-wise MSE:

$$\text{BackgroundStability} = \text{RobustMax}_t \left(\text{MSE} \left(\text{BG}(I^t), \text{BG}(I^0) \right) \right). \tag{1}$$

For every frame, only background pixels shared between both images are considered. The shared background is computed as the intersection between the background mask of both frames. In this case, we use the robust maximum operation to focus on frames where the background is heavily altered. In order to scale this metric into the $[0, 1]$ range, we apply decaying exponential scaling $S' = \exp[-50 \cdot S]$.

Motion similarity. The introduced metric to evaluate motion is based on the motion encoder from DisMo (Ressler-Antal et al., 2025). This model encodes motion representations that are independent from appearance or object identity, while retaining the surrounding scene context of the queried region. This way, we are able to disentangle motion from visual properties, without relying on simplistic metrics, such as comparing object centroid position. We compute the metric by taking the robust minimum over the video of the cosine similarity of the embeddings from the generated video and each different rollout, then select the best match across rollouts,

$$\text{MotionSimilarity} = \max_r \left(\text{RobustMin}_t \left(\langle \text{Dismo}(I^t) \cdot \text{Dismo}(\tilde{I}_r^t) \rangle \right) \right), \tag{2}$$

where $\tilde{I}_r$ indicates the different r reference frames. Before the embeddings are computed, the sequences are temporally aligned.

Appearance stability. This metric quantifies semantic distortions of objects across the generated videos. For that, we compare all generated frames with the first context frame, which matches the original render. For a given frame I^t, we mask each object (I_i^t) and extract its CLS token using DINOv2. Then, we compute the cosine similarity between the tokens from the first and subsequent frames. The stability metric is computed by taking our robust minimum across the video and across objects. This way, the metric is sensitive to object alterations which happen only in a few frames or only happen to a single object. We compute

$$\text{AppearanceStability} = \text{RobustMin}_{i,t} \left(\langle \text{CLS}(I_i^t) \cdot \text{CLS}(I_i^0) \rangle \right), \tag{3}$$

where $\langle \cdot \rangle$ denotes the cosine similarity.

3D shape stability. Object shape is usually overlooked or indirectly measured in other benchmarks. We take advantage of advances in 3D shape estimation models to design a novel metric to quantify object morphing in video models. We run SAM3D on the generated videos by prompting the model with the object segmentation masks predicted by SAM3, and obtain object meshes M_i^t. For each video, we align all meshes to the reference by optimizing scale and rotation. For each object i, we compute the Chamfer Distance (CD) between predicted meshes for the initial and all subsequent frames,

$$\text{3DShapeStability} = \text{RobustMax}_{i,t} \left(\text{CD}(M_i^t, M_i^0) \right). \tag{4}$$

We again apply the exponential scaling used for background stability to bring CD into the $[0, 1]$ range.

Physical plausibility. This metric captures high-level information about the physical events in the video, covering failure modes that are hard to detect automatically with the low-level metrics above. We design 5 general questions, as well as 5 templates for each physical event. The templates are individually adjusted for each video by substituting sequence-specific details, such as object classes. The prompt is:

VLM-as-judge input prompt

You are evaluating a short video of a physics simulation. An object in the scene was pushed just before the video starts. Watch the video carefully and answer all of the following questions about what you observe. For each question provide a True or False answer. Provide brief comments justifying your answers and rate your overall confidence from 0.0 (not confident at all) to 1.0 (fully confident).

Video description: <video_description>

Questions: <video_questions>

Respond ONLY with valid JSON using exactly this structure (no extra text outside the JSON):

```
{
"answers": {
  "question_key1": <answer1>,
  ...
},
"comments": "<brief justification for each answer>",
"confidence": <float 0.0-1.0>
}
```

And the specific questions can be found in Table 5. The final metric is computed by counting the questions whose answer does not match the expected outcome, and scaling with the inverse of that count:

$$\text{PhysicalPlausibility} = \left(1 + \sum_{b} \mathbf{1}[b_{\text{VLM}} \neq b_{\text{GT}}]\right)^{-1}. \tag{5}$$

This way, even a few negative answers result in low physical plausibility, whereas a linear metric would still assign such a video a high score.

Success rate. We compress the quality of each video by calibrating thresholds in all our metrics. This way, we prevent catastrophic failures in one metric from being smoothed by other results. The thresholds are calibrated using the human study. We consider successful videos in an annotated metric as those which achieve a median score of 3 or higher. We then calibrate thresholds on the corresponding metrics by minimizing the absolute difference between false positive rate and false negative rate.

We study the stability of the thresholds by sweeping them individually and computing the interval of thresholds where the success rate ranking is preserved (Table 6). The wide ranges show the stability of the selected thresholds. We also bootstrap the subset of 480 human-study videos 1,000 times and use the resamples to compute the confidence interval (CI) of each threshold (Table 6), which always lies within the ranking-stable interval. We therefore conclude that the thresholds are stable and the rankings are not sensitive to noise in the human study.

Sensitivity to interventions. For each intervention type, we group videos that differ only along that intervention and were generated with the same seed. For a group g and metric m, the raw sensitivity $d_m(g)$ is the mean absolute pairwise difference of m between the videos in the group. To make metrics comparable, we normalize by the seed noise σ_m, the mean absolute pairwise difference of m between seeds of the same input, averaged over all sequences and models. The sensitivity of a model to an intervention is

$$\text{Sensitivity} = \frac{1}{|M|} \sum_{m \in M} \frac{1}{|G|} \sum_{g \in G} \frac{d_m(g)}{\sigma_m}, \tag{6}$$

Table 5: **Questions for physical plausibility metric.**

Event	Questions
Fall	Does the {object_name} fall off the {surface} when it reaches the edge? Does the {object_name} hit the ground? Does the {object_name} change its direction while falling? Does the {object_name} move on an arc-shaped path while falling? Does the {object_name} accelerate while falling?
Collision	Does the {object_name} contact with the {collision_target}? Does the {object_name} come to a stop before contacting the {collision_target}? Is the motion of {object_name} affected by the collision with the {collision_target}? Is the reaction to the impact realistic considering the size and mass of the objects? Are objects deformed or broken during or after the collision?
Occlusion	Does the {object_name} move behind the {occluding_object} during the video? Does the {object_name} reappear on the other side after being occluded by the {occluding_object}? Does the {object_name} move on a straight path during the whole video? Does the {object_name} disappear after being occluded by the {occluding_object}? Does the {object_name} change its appearance after being occluded by the {occluding_object}?
Shared questions	Is the background static throughout the video? Does the {object_name} maintain its color and shape throughout the video? Do new objects appear on the scene during the video? Do objects move smoothly without sudden jumps or teleportation during the video? Do objects move without forces acting on them?

Table 6: Statistical study of success thresholds for all parameters. (*) Since the physical plausibility metric is discrete, the whole range between $1/3$ and $1/2$ results in the same outcome.

Metric	Calibrated threshold	Ranking-stable interval	Calibration 95% CI
Background Stability	0.30	$[0.08, 0.42]$	$[0.21, 0.34]$
Appearance Stability	0.48	$[0.16, 0.62]$	$[0.45, 0.51]$
3D Shape Stability	0.60	$[0.00, 0.74]$	$[0.55, 0.64]$
Physical Plausibility	0.49	$[0.21, 0.50]$	$[0.33, 0.50]^*$
Motion Similarity	0.57	$[0.55, 0.76]$	$[0.54, 0.59]$

where M is the set of metrics and G the set of groups for that intervention across all seeds. The average sensitivity in Table 9 is the mean over the interventions available for each event.

Since σ_m is averaged over all models, it is a single constant per metric that is shared by every model. It therefore only rescales each metric and does not depend on a model's own seed variability. A sensitivity of 1 means that an intervention changes quality as much as resampling the seed typically does.

Note that sensitivity measures the consistency of generation *quality*, not the similarity between generated videos. Since each video is evaluated against the references of its own input, two videos with different trajectories, e.g., after an object intervention changes the object's mass, obtain identical scores if both match their respective references.

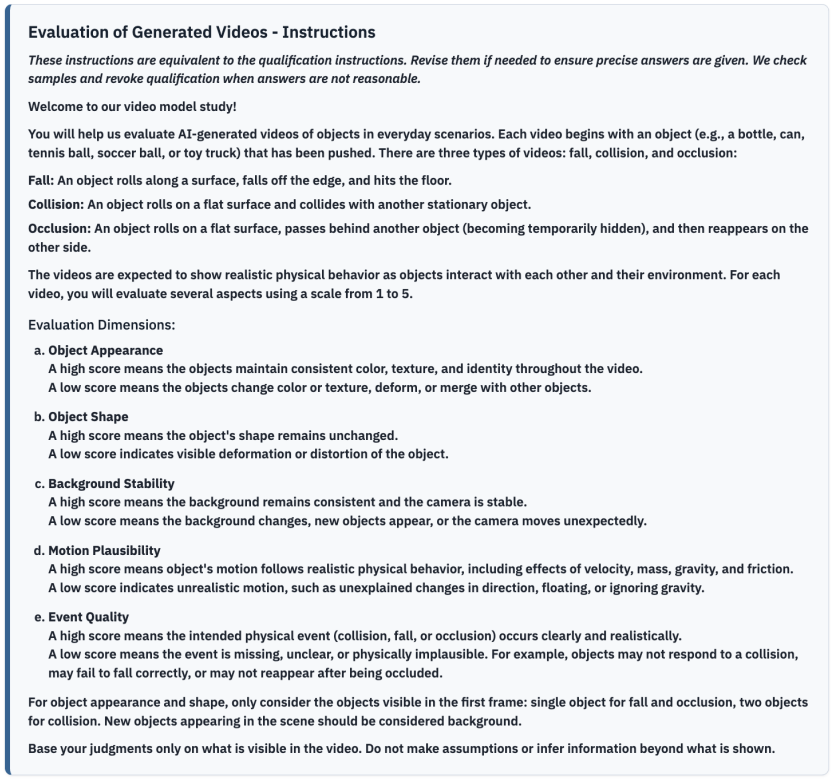

Figure 6: Instructions of the human study.

D HUMAN STUDY

We present the instructions of the human study in Figure 6 and the GUI for annotating videos along the considered quality axes in Figure 7. All annotators received detailed instructions and needed to pass a qualification exam before participation. The 8 hired Prolific annotators received a compensation of £14 per hour.

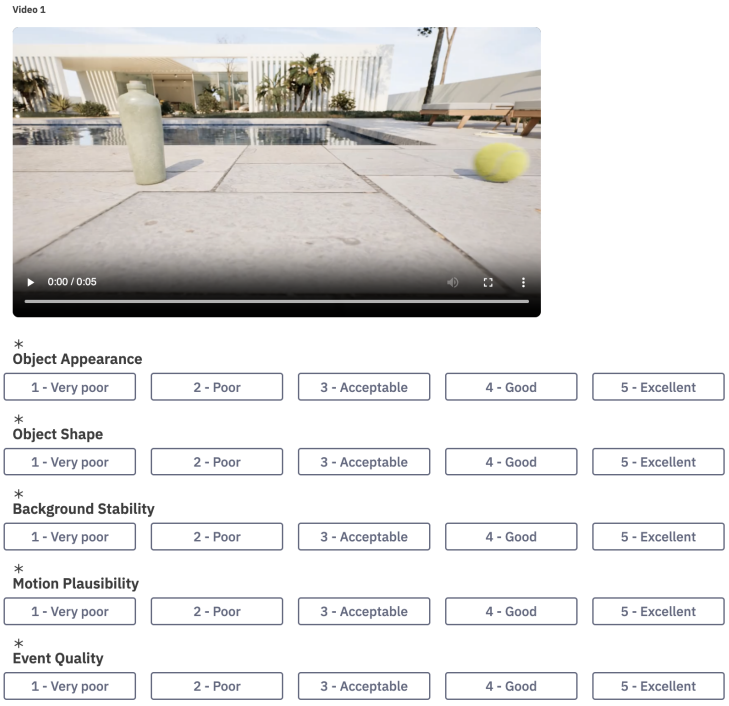

Figure 7: GUI used in the user study for annotating video quality.

E ADDITIONAL EXAMPLES

In Figures 8 to 10 we show rollout examples of generated videos.



Figure 8: **Additional generated examples: collision.**

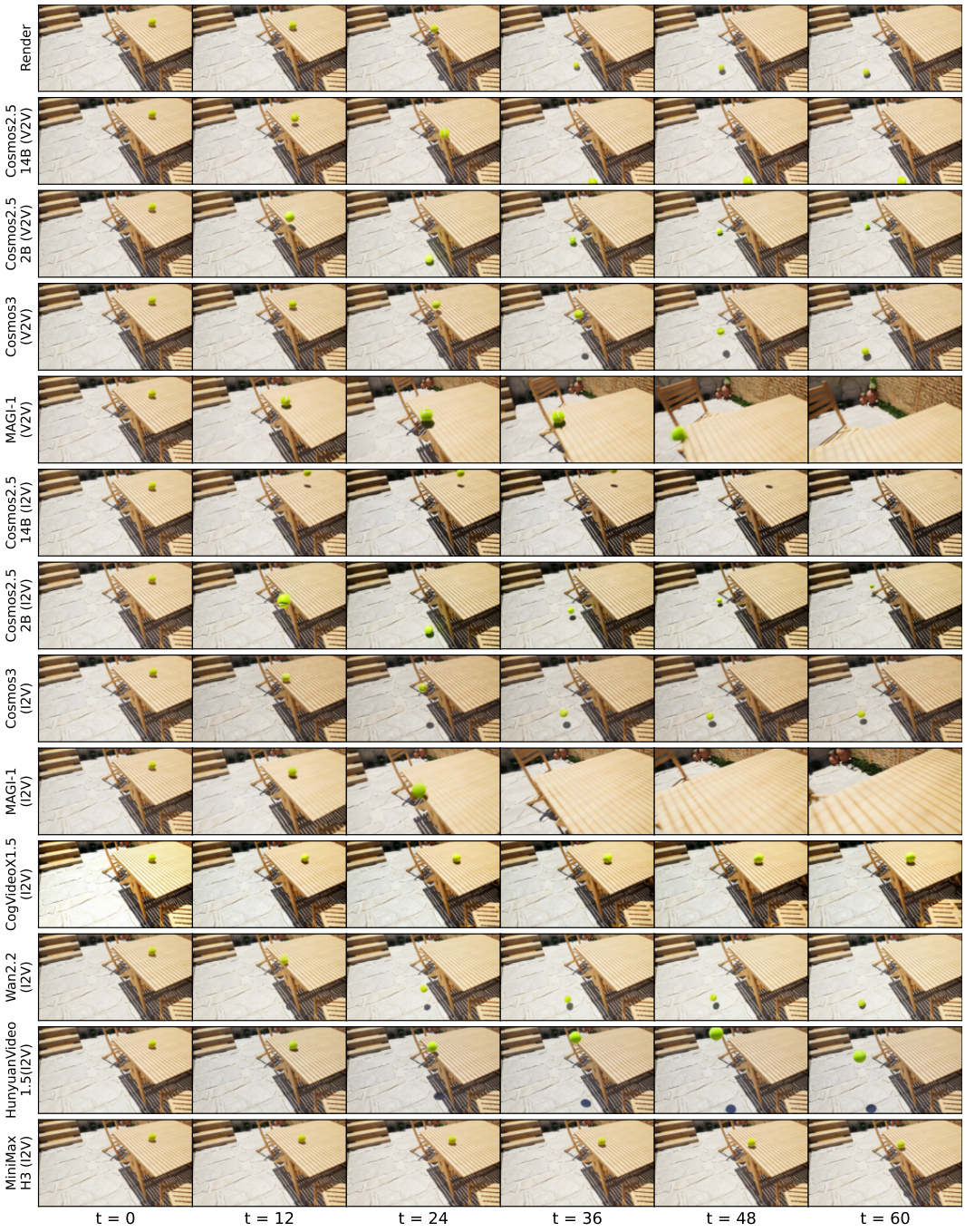

Figure 9: **Additional generated examples: fall.**

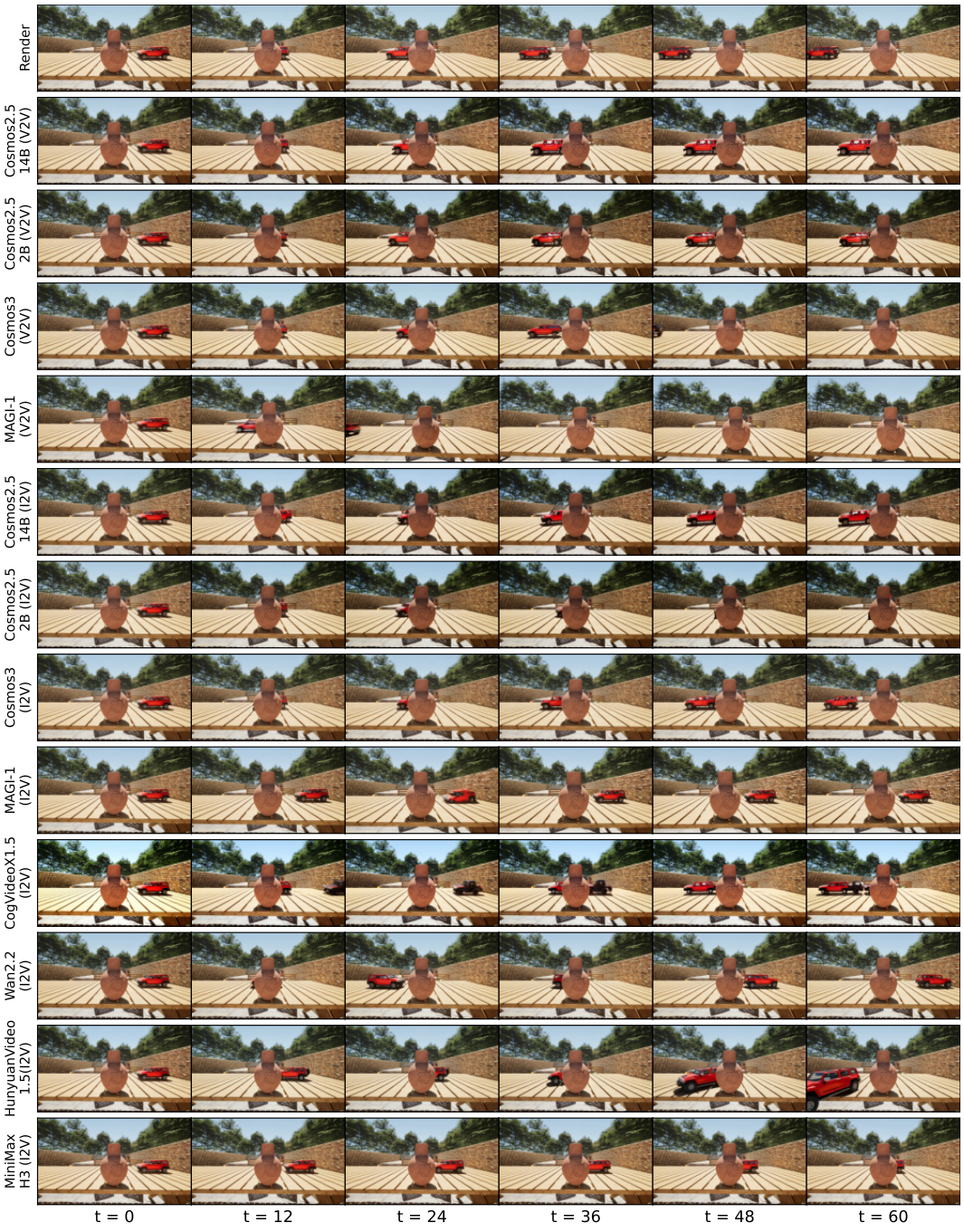

Figure 10: **Additional generated examples: occlusion.**

F REAL-DATA EXPERIMENT

We conduct a small-scale experiment to estimate the simulation-to-real gap on our benchmark. We imitate sequences on the benchmark by recording experiments on a real environment. In particular, we run all three physical events, using three different objects. After that, we generate videos for all models following the benchmark's prompt structure in I2V setting, for a total of 315 generated videos.

The numerical results (Table 7) show similar results as for the main benchmark metrics, with MiniMax-H3 and Cosmos H3 leading in all metrics and MAGI-1 and CogVideo performing worst. In general, all metrics stay in similar ranges as for synthetic data. The qualitative results in Figure 11 also show similar failure modes in both data regimes. These experiments suggest that the simulation-to-real gap is small.

Table 7: **Benchmark metrics on real videos.**

Mode	Model	Bg Stab.	Motion Sim.	App. Stab.	3D Shape Stab.	Physical Plaus.	Success Rate
I2V	Cosmos2.5-2B	0.44	0.42	0.54	0.74	0.52	0.04
	Cosmos3-Super	0.64	0.60	0.58	0.76	**0.61**	0.09
	Magi-1-4.5B	0.25	0.49	0.65	0.81	0.41	0.00
	CogVideoX1.5-5B	0.21	0.37	0.57	0.70	0.43	0.00
	Wan2.2-14B	0.58	0.56	0.60	0.79	0.60	0.07
	HunyuanVideo-1.5	0.51	0.46	0.56	0.70	0.55	0.02
	MiniMax H3	**0.73**	**0.61**	**0.67**	**0.86**	0.58	**0.16**

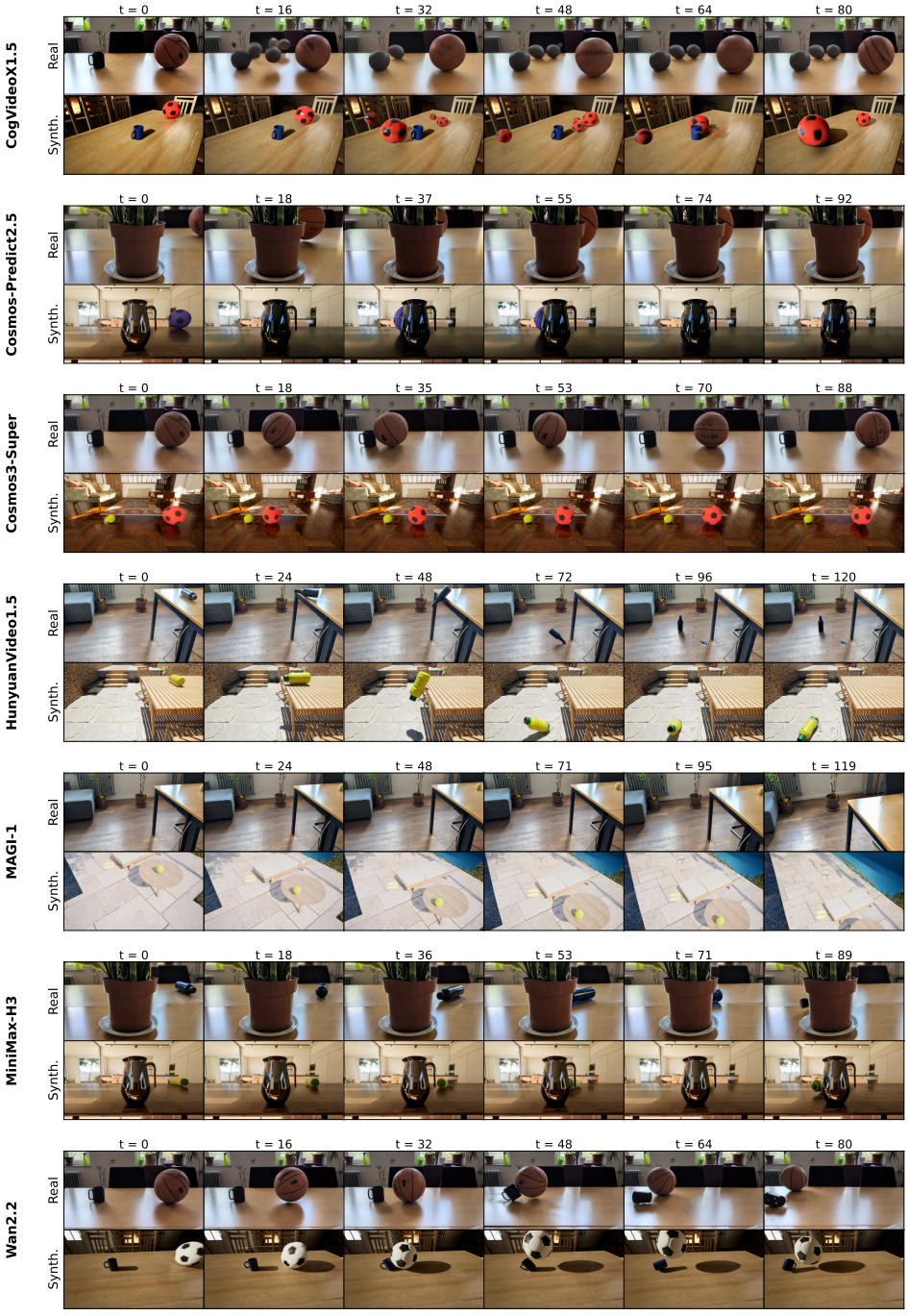

Figure 11: **Real and synthetic data rollout comparison.** Each model shows a different common failure mode of the models. In order: severe object distortion and duplication, unexplained stop, unexplained back motion, object distorting and reversing sides, undesired camera motion, unexplained rotation and incorrect physical properties. The fact that similar failure modes are observed in both data regimes suggest a small gap in the abilities the benchmark quantifies.

G Detailed Results

We provide benchmark metrics for each physical event in Table 8 and sensitivity to interventions in Table 9. We also provide the failure rates per metric in Figure 12.

Table 8: **Benchmark metrics per physical event.** Metrics are normalized to $[0, 1]$ and higher is better. Best, second-, and third-best values per metric are indicated by bold, underline, and dashed underline, respectively. Values are ranked within each event, across both conditioning modes.

Mode	Model	Bg Stab.	Motion Sim.	App. Stab.	3D Shape Stab.	Physical Plaus.	Success Rate
Fall							
V2V	Cosmos2.5-2B	0.80	0.64	0.52	0.62	0.47	0.16
	Cosmos2.5-14B	0.59	0.55	0.49	0.56	0.43	0.10
	Cosmos3-Super	0.82	0.67	0.54	0.68	0.87	0.27
	Magi-1-4.5B	0.23	0.48	0.47	0.55	0.38	0.00
I2V	Cosmos2.5-2B	0.64	0.54	0.48	0.56	0.45	0.08
	Cosmos2.5-14B	0.57	0.50	0.47	0.55	0.43	0.06
	Cosmos3-Super	0.67	0.63	0.52	0.70	**0.88**	0.19
	Magi-1-4.5B	0.22	0.47	0.52	0.55	0.39	0.02
	CogVideoX1.5-5B	0.41	0.37	0.42	0.49	0.41	0.01
	Wan2.2-14B	**0.83**	**0.68**	0.56	0.73	0.47	0.24
	HunyuanVideo-1.5	0.76	0.58	0.55	0.62	0.42	0.09
	MiniMax H3	0.62	0.53	**0.66**	**0.75**	0.84	**0.40**
Collision							
V2V	Cosmos2.5-2B	**0.79**	0.65	0.54	**0.77**	0.90	**0.31**
	Cosmos2.5-14B	0.51	0.56	0.52	0.76	0.86	0.18
	Cosmos3-Super	0.78	**0.67**	0.52	0.73	**0.94**	**0.31**
	Magi-1-4.5B	0.24	0.45	0.45	0.61	0.59	0.02
I2V	Cosmos2.5-2B	0.57	0.58	0.51	0.70	0.80	0.17
	Cosmos2.5-14B	0.46	0.52	0.50	0.69	0.85	0.10
	Cosmos3-Super	0.72	0.61	0.48	0.70	0.93	0.16
	Magi-1-4.5B	0.24	0.44	0.49	0.57	0.65	0.02
	CogVideoX1.5-5B	0.36	0.30	0.38	0.47	0.65	0.01
	Wan2.2-14B	0.67	0.53	0.50	0.75	**0.94**	0.17
	HunyuanVideo-1.5	0.66	0.42	0.48	0.59	0.42	0.02
	MiniMax H3	0.54	0.53	**0.57**	0.73	**0.94**	0.28
Occlusion							
V2V	Cosmos2.5-2B	0.59	0.64	0.52	0.54	0.93	0.19
	Cosmos2.5-14B	0.55	0.60	0.56	0.50	0.88	0.14
	Cosmos3-Super	0.77	**0.73**	0.58	0.72	**1.00**	0.35
	Magi-1-4.5B	0.11	0.52	0.55	0.59	0.74	0.02
I2V	Cosmos2.5-2B	0.59	0.61	0.52	0.61	0.92	0.15
	Cosmos2.5-14B	0.56	0.58	0.51	0.51	0.89	0.14
	Cosmos3-Super	0.82	0.69	0.54	0.70	**1.00**	0.31
	Magi-1-4.5B	0.10	0.50	0.59	0.62	0.74	0.05
	CogVideoX1.5-5B	0.43	0.48	0.44	0.55	0.93	0.08
	Wan2.2-14B	0.84	0.58	0.58	0.63	0.97	0.22
	HunyuanVideo-1.5	0.55	0.49	0.54	0.60	0.95	0.12
	MiniMax H3	**0.85**	0.72	**0.69**	**0.82**	0.99	**0.49**

Table 9: **Benchmark sensitivities per physical event.** Lower is better: a lower value indicates that generation quality changes less under that intervention. Best, second-, and third-best values per metric are indicated by bold, underline, and dashed underline, respectively. Values are ranked within each event, across both conditioning modes. Viewpoint is not varied for occlusion events, so those entries are marked "–".

Mode	Model	Scene	Object	Appearance	View	Average
Fall						
V2V	Cosmos2.5-2B	1.14	1.17	0.73	1.09	1.03
	Cosmos2.5-14B	1.55	1.27	0.75	1.28	1.21
	Cosmos3-Super	0.90	**1.08**	0.53	**0.85**	**0.84**
	Magi-1-4.5B	1.36	1.36	1.00	1.41	1.28
I2V	Cosmos2.5-2B	1.34	1.35	0.83	1.33	1.21
	Cosmos2.5-14B	1.53	1.33	0.79	1.19	1.21
	Cosmos3-Super	1.06	1.14	0.57	1.10	0.97
	Magi-1-4.5B	1.42	1.34	1.04	1.50	1.33
	CogVideoX1.5-5B	1.25	1.32	1.04	1.32	1.23
	Wan2.2-14B	**0.89**	1.18	0.54	0.87	0.87
	HunyuanVideo-1.5	1.12	1.41	0.75	1.10	1.09
	MiniMax H3	1.47	1.09	**0.48**	1.24	1.07
Collision						
V2V	Cosmos2.5-2B	1.18	1.19	0.68	1.12	1.04
	Cosmos2.5-14B	1.50	1.37	0.78	1.42	1.26
	Cosmos3-Super	**1.07**	1.05	0.60	**1.08**	**0.95**
	Magi-1-4.5B	1.62	1.46	1.06	1.65	1.45
I2V	Cosmos2.5-2B	1.57	1.48	0.93	1.53	1.38
	Cosmos2.5-14B	1.55	1.45	0.89	1.42	1.33
	Cosmos3-Super	1.19	1.18	0.67	1.13	1.04
	Magi-1-4.5B	1.81	1.48	1.16	1.71	1.54
	CogVideoX1.5-5B	1.59	1.52	1.25	1.53	1.47
	Wan2.2-14B	1.25	1.19	0.69	1.14	1.07
	HunyuanVideo-1.5	1.26	1.25	0.82	1.17	1.13
	MiniMax H3	1.48	**1.04**	**0.57**	1.52	1.15
Occlusion						
V2V	Cosmos2.5-2B	1.13	1.18	0.72	–	1.01
	Cosmos2.5-14B	1.35	1.18	0.77	–	1.10
	Cosmos3-Super	0.75	**0.81**	0.49	–	0.69
	Magi-1-4.5B	1.33	1.28	1.04	–	1.22
I2V	Cosmos2.5-2B	1.21	1.09	0.77	–	1.02
	Cosmos2.5-14B	1.36	1.23	0.88	–	1.16
	Cosmos3-Super	0.70	0.88	0.50	–	0.69
	Magi-1-4.5B	1.45	1.36	1.12	–	1.31
	CogVideoX1.5-5B	1.22	1.33	1.07	–	1.21
	Wan2.2-14B	0.82	1.11	0.61	–	0.85
	HunyuanVideo-1.5	1.22	1.41	0.84	–	1.16
	MiniMax H3	**0.63**	**0.81**	**0.31**	–	**0.58**

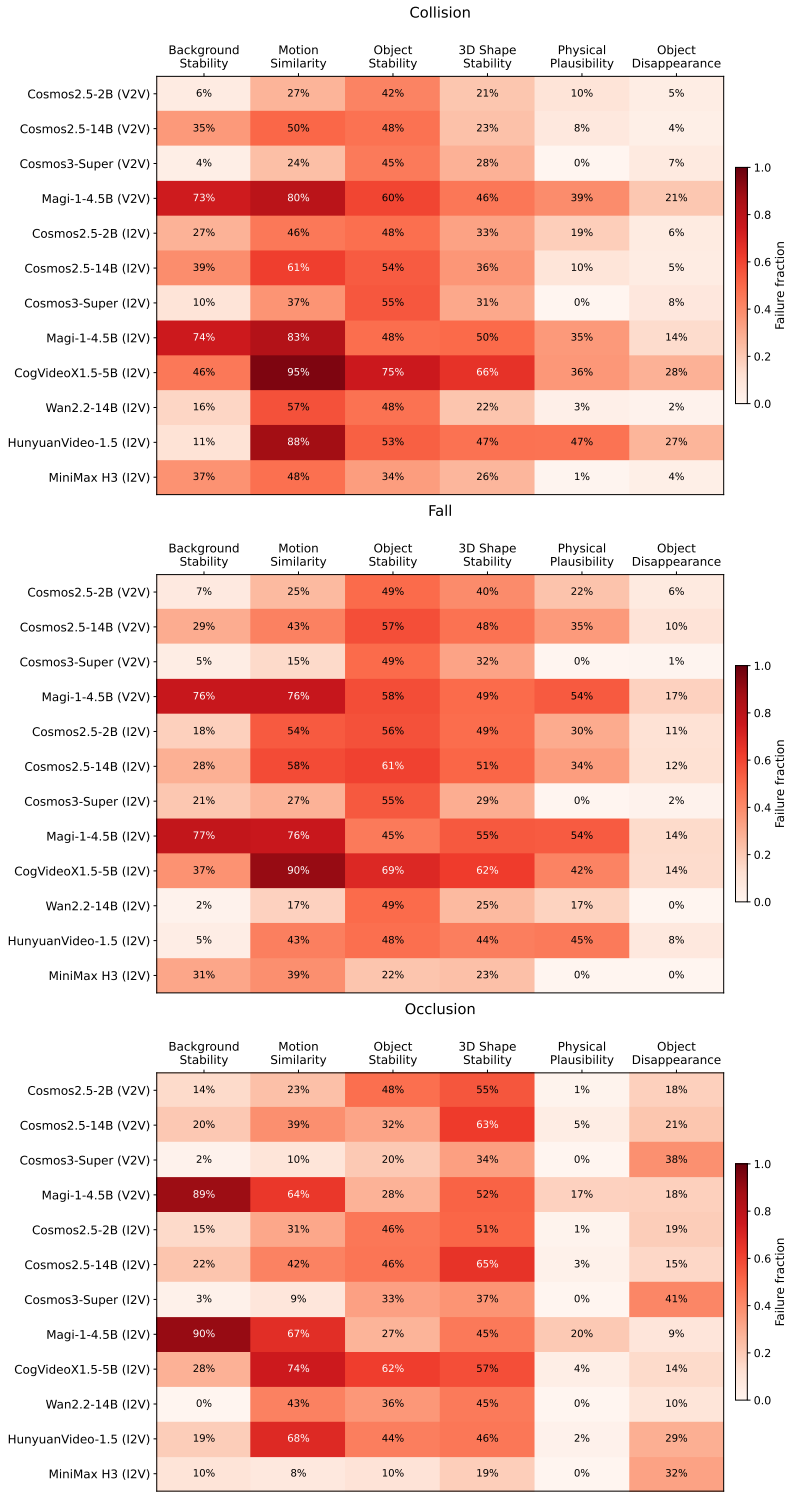

Figure 12: **Failure rates per metric.** For each event, we provide the fraction of videos failing in each metric.

Table 10: **Human study results.** Average median human annotation score per model and metric, 1–5 scale, higher is better. Best, second-, and third-best values per metric are indicated by bold, underline, and dashed underline, respectively.

Mode	Model	Object Appear.	Object Shape	Bg Stab.	Motion Plaus.	Event Quality
V2V	Cosmos2.5-2B	3.65	3.30	**4.44**	1.62	1.88
	Cosmos2.5-14B	3.70	3.33	3.88	1.76	1.94
	MAGI-1	2.57	2.57	2.26	1.40	1.33
I2V	Cosmos2.5-2B	3.42	3.29	3.92	1.53	1.68
	Cosmos2.5-14B	3.47	3.14	3.83	1.76	1.91
	MAGI-1	2.62	2.70	2.20	1.38	1.29
	CogVideoX1.5	1.59	1.37	2.78	1.08	1.16
	Wan2.2	**4.15**	**4.03**	4.21	**2.33**	**2.68**

H Asset Licenses and Release Documentation

CRONOS uses third-party Unreal Engine scene assets to construct the rendered simulation environments. All 3D assets were purchased under Epic Content License Agreement (ECLA), and selected only when authors allowed GenAI-related research and benchmark generation. We use the assets as part of controlled synthetic scenes rendered in Unreal Engine; the benchmark does not rely on scraped videos or unlicensed real-world footage.

We release the generated benchmark videos and evaluation code. The release includes the rendered RGB videos, text prompts, benchmark metadata describing the physical event and intervention values for each video, and the code used to compute the evaluation metrics reported in the paper.

We do not redistribute third-party Unreal Engine source assets whose licenses restrict redistribution. Instead, released artifacts are be limited to the generated benchmark data, code, and metadata that can be shared under the applicable licenses. The released package will include license information for the dataset and code, along with attribution and license notes for third-party tools and assets used to construct the benchmark.