WORLD BEFORE PIXELS: AGENTIC PHYSICAL SIMULATION FOR TEXT-TO-VIDEO SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative text-to-video synthesis has achieved striking visual realism, yet current models still struggle to ensure that scene geometry supports the requested interaction and that the resulting motion follows physical laws. Faithfully modeling the world requires understanding not only visual appearance, but also object geometry and material properties, physical states and interactions, and the causal dynamics that govern how these states evolve—knowledge that is difficult to acquire implicitly from even massive collections of video data. We present an agentic framework that addresses these challenges by constructing, validating, and simulating an explicit physical world before synthesizing its appearance. Given a text prompt, our agent constructs a structured and executable scene whose geometry, materials, physical roles, and simulation parameters expose the variables governing the requested event. Simulation then turns these coupled choices into observable consequences, including rigid-body contact, deformation, and fluid motion, enabling a vision-language controller to assess the simulated outcome and revise the scene through validated edits and re-execution. This closed loop progressively aligns the constructed world with the intended event. Finally, as the requests from the initial text prompt are met, verified geometric priors are rendered from the simulated rollouts, supplying aligned depth, contour, and material controls for a pretrained video model. This combines simulated dynamics with generative appearance synthesis without additional model training. Experiments across diverse physical scenarios show improved physical fidelity over direct text-to-video generation and ablated variants.

1 INTRODUCTION

Text-to-video synthesis has achieved striking visual realism, producing detailed scenes and compelling motion from natural-language descriptions (Ho et al., 2022b; Blattmann et al., 2023). Yet faithfully depicting a physical event requires more than plausible appearance: the scene geometry must support the requested interaction, and the resulting motion must follow physical laws. For liquid to flow into a vessel, the container needs an accessible cavity, the source must be positioned appropriately, and the liquid must respond to gravity and contact. Convincing objects and textures alone do not establish that this event occurred. Benchmarks of generated videos document such gaps between visual quality and physical commonsense (Bansal et al., 2025; Meng et al., 2025).

Modeling such events requires geometry and material properties, physical states and interactive dynamics, many of which are only partially observable in images. Learning their joint effects implicitly from video therefore remains challenging even at scale. Text-to-video models must infer these physical conditions from the prompt using their learned video priors. Even when generation is additionally conditioned on a first frame, the specified initial appearance does not fully determine collision geometry, material response, or the forces governing subsequent motion. Essential physical decisions thus remain implicit despite the additional visual guidance. This motivates constructing an explicit physical world and computing its evolution before synthesizing appearance.

Physics simulation provides a natural way to compute the resulting motion from explicit geometry, material properties, and initial and boundary conditions. Prior work demonstrates its value for rigid-body animation and fluid synthesis (Liu et al., 2025; Chen et al., 2025b). However, even though a simulator faithfully evolves the world it is given, this does not guarantee that the intended event

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.

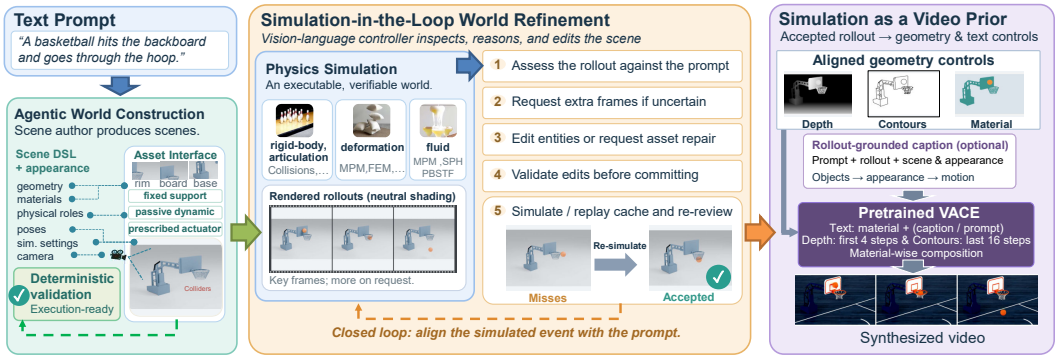


Figure 1: **Overview.** Our framework combines agentic world construction, simulation-in-the-loop refinement, and video synthesis guided by the accepted rollout.

will occur. Thus, successful execution alone is insufficient: the challenge is not only to simulate a world, but to construct one that is executable, verifiable, and readily editable in response to desired consequences, all in an efficient manner. Recent agentic approaches show that execution feedback can help revise generated scenes (Li et al., 2026; Zhang et al., 2026), motivating an editable world whose geometry and physical parameters can be refined through rollout inspection.

We present an agentic framework that constructs, simulates, and refines a physical world before generating its appearance. Given a text prompt, a scene author instantiates a structured scene description specifying geometry, materials, physical roles, and simulation settings. Asset interfaces expose interaction geometry, such as cavities and outlets, while the scene representation distinguishes fixed supports, passive objects, and prescribed actuators. Deterministic validation checks the scene before a simulator computes rigid-body interactions, deformation, or fluid motion. A vision-language controller inspects the rendered rollout, requests additional frames to resolve ambiguous interactions, and proposes edits to named scene entities and parameters. Each revision must pass validation and be re-executed, forming a closed loop that progressively aligns the constructed world with the intended event. Once a rollout is accepted, its cached geometry supplies aligned depth, contour, and material controls from the same sequence used for inspection. Following Fluid Composer (Chen et al., 2025b), staged geometric conditioning and material-wise composition guide pretrained VACE (Jiang et al., 2025) to synthesize appearance without additional model training. Across diverse physical scenarios, our framework achieves higher final-video physical fidelity than its ablated variants and direct text-guided video generation baselines, even when these baselines are additionally conditioned on a first frame. Our contributions are:

- An agentic framework that translates text into an editable physical world through a structured, validated scene representation of interaction geometry, materials, and physical roles;
- A simulation-in-the-loop refinement mechanism combining visual inspection, additional frame requests, and validated scene revisions followed by renewed execution and review;
- An integration of reviewed simulation geometry with pretrained video synthesis, evaluated across diverse physical scenarios against direct generation baselines and ablated variants.

2 Related Work

Text-to-Video Synthesis. Modern text-to-video models have rapidly advanced through diffusion-based generation, latent video representations, and transformer architectures (Ho et al., 2022b;a; Singer et al., 2023; Guo et al., 2024; Blattmann et al., 2023; Bar-Tal et al., 2024; Yang et al., 2025; Wan Team et al., 2025). Structured conditioning further improves controllability through spatial, temporal, camera, and geometric guidance (Zhang et al., 2023; Wang et al., 2023; He et al., 2025; Gu et al., 2025b; Jiang et al., 2025). These models provide strong appearance and motion priors, but typically operate on visual conditions rather than an explicit, editable physical world. Our framework uses pretrained VACE (Jiang et al., 2025) as the appearance backend while deriving its geometric controls from a simulated scene whose geometry, materials, and physical parameters can be explicitly inspected and revised.

Physics-Grounded Video Generation. A growing body of work incorporates physical simulation or learned dynamics into video generation. Explicit simulation has been used to guide rigid-body, deformable-body, and fluid animation (Lv et al., 2024; Liu et al., 2025; Chen et al., 2025a; Xie et al., 2025; Liu et al., 2026; Li et al., 2025; Chen et al., 2025b), while complementary approaches learn force-conditioned motion, material response, or physics-aware dynamics directly from data (Gillman et al., 2025; 2026; Wang et al., 2025; Yuan et al., 2026; Pathak et al., 2026). These methods demonstrate that physical structure can substantially improve generated motion. Our approach places the simulator inside an iterative world-construction loop: rather than using simulation only as a source of motion guidance, the simulated outcome is inspected to revise the underlying scene until it realizes the intended event. The accepted simulation then provides temporally aligned geometric and material controls for generative appearance synthesis, following the conditioning strategy of Fluid Composer (Chen et al., 2025b).

Agentic World Building. Recent agentic systems use executable programs, rendering, and iterative feedback to construct and refine visual worlds (Yao et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Hu et al., 2024; Lu et al., 2025; Wang et al., 2026; Gu et al., 2025a; Yin et al., 2026). More closely related, PhysAgent (Li et al., 2026) revises physical programs using simulation feedback, while GS-Agent (Zhang et al., 2026) coordinates specialized agents for constructing physical scenes from text. Our framework builds on this simulation-feedback paradigm with a structured physical-world representation designed for iterative execution: interaction geometry, materials, physical roles, and simulation parameters are exposed as named, editable scene components; deterministic validation and rollout inspection guide revisions; and the same accepted scene directly supplies the controls used for final video synthesis. This creates a closed loop from text, to an executable physical world, to verified dynamics, and finally to photorealistic video.

3 Agentic Physical World Building

3.1 Overview

The framework proceeds through physical-world construction, rollout-based refinement, and rollout-conditioned video synthesis. Given an event description x, a scene author constructs a typed physical program and appearance specification (section 3.2). A controller executes the program, while a rollout reviewer inspects the resulting frames, requests further evidence when needed, and accepts or revises the scene (section 3.3). After acceptance, a caption pipeline can optionally describe the observed event, while cached rollout geometry supplies the video controls (section 3.4). A shared structured document keeps geometry, dynamics, camera, and appearance jointly editable across construction and refinement.

3.2 Agentic World Construction

A representation for physical decisions. The author produces $\mathcal{D} = (S, A)$, with a declarative Scene DSL S and appearance specification A:

$$S = (\theta_{\text{sim}}, \mathcal{O}, \mathcal{E}, \mathcal{K}, \theta_{\text{out}}), \qquad o_i = (g_i, m_i, T_i, u_i, a_i), \tag{1}$$

Here θ_{sim} contains gravity, timestep, duration, domain, and solver settings; $\mathcal{O}$, $\mathcal{E}$, and $\mathcal{K}$ specify objects, liquid emitters, and cameras; and θ_{out} controls output sampling. An object has geometry g_i, material m_i, pose T_i, motion u_i, and asset reference and metadata a_i. Asset interfaces describe parts, joints, allowed motion, and material response; scene instances specify placement, initial state, and control (Appendix A.1). Stable identifiers address entities and fields, with lengths in meters and times in seconds.

The author distinguishes fixed supports, passive dynamic objects, and prescribed actuators: a tool may follow a trajectory while contact drives its target's deformation. Separately, A assigns each object and emitter to one appearance group with a text prompt and transparency flag, plus a background group. Physical parameters specify response; text specifies visible surfaces.

Geometry that supports the interaction. The asset interface takes required geometry, interaction, dimensions, and material behavior. For non-articulated objects, bounded repository search

precedes procedural aDSL generation (Wang et al., 2026), with Blender for format conversion. Articulated designs instead compile deterministically to URDF; visual inspection of offscreen renders and simulator probes gate registration and trigger revisions (Appendix A.1). The mechanism design is proposed by a model; failed visual or compatibility checks feed back to this design before registration. A non-articulated asset is unavailable only after both retrieval and procedural generation fail, while articulated acquisition depends on its dedicated route. Simple scene elements, including ground planes and spherical projectiles, use primitives.

Interfaces expose metric bounds, materials, scaling, collision geometry, support surfaces, and source anchors. Fixed assets share visual and collision coordinates but may omit decorative colliders or preserve vessel cavities. Outlet-bound emitters must match transformed position, direction, and aperture shape, with nozzles no larger than the outlet. Elastic assets require volumetric geometry, effective stiffness, Poisson's ratio, and density. Geometry checks and simulator probes gate registration; rollouts test the resulting interaction. For vessels, these checks cover both recognizable shape and an open collision cavity that supports the intended interaction.

Validation before execution. A deterministic validator checks types, finite values, parameter ranges, identifiers, references, material–geometry compatibility, appearance coverage, and asset-interface compliance. Errors prompt bounded document corrections; contact, containment, and event success require simulation.

Once the document passes these representation-level checks, it is compiled to a backend selected for the requested physical phenomenon. The validated DSL compiles to Genesis (Genesis Authors, 2024), supporting rigid and articulated bodies, MPM elasticity with separable per-body contact, MPM/SPH liquids, DEM sand, impulse-released precomputed fracture fragments, and flexible ropes represented by articulated chains with continuous rendered surfaces. The local PBSTF extension supports one liquid family with fixed rigid colliders and rejects dynamic rigid coupling. Scene-selected solvers must satisfy their backend constraints (Appendix A).

For MPM liquids, scene geometry determines simulation and reconstruction scales before execution; geometry-changing revisions recompute them, while camera-only edits reuse them. Appendix A gives the scale rule and defaults.

3.3 Simulation-in-the-Loop World Refinement

Observing the requested event. A rollout is accepted only when the reviewer observes the requested initiating cause, decisive interaction, and visible aftermath, together with the deterministic validation and numerical checks described in Appendix A. At iteration k, the controller simulates $\mathcal{D}_k = (S_k, A_k)$ and renders captured geometry as preview V_k with neutral, opaque inspection shading. The reviewer receives x, the document with resolved cameras, video metadata, and registered interfaces in world coordinates. It initially sees eight evenly spaced frames and may request indexed frames, including the full sequence, to resolve ambiguous interactions. Up to six additional requests append frames and temporal indices to the same conversation. Each new simulation starts a fresh visual session. Requests can focus on the interval around impact to distinguish a missed collision from brief contact; earlier observed frames remain available in the same conversation.

Revising the scene from visual evidence. The reviewer explains an acceptance, scene revision, asset repair, frame request, or unresolved outcome. Revisions specify edits Δ_k and optional appearance replacement A'_k; repair reacquires defective assets. Unresolved outcomes must cite unsupported capabilities or missing input. Alongside frames, the reviewer receives floor-crossing checks, articulation degrees of freedom, and geometric measurements (Appendix A). The same visual session identifies the failure and specifies edits: a stream missing its vessel can motivate a placement change, while an occluded interaction can motivate a camera change.

Edits change, add, or remove named entities on a document copy and commit only when the complete result passes validation:

$$\mathcal{D}_{k+1} = \mathcal{P}(\mathcal{U}(\mathcal{D}_k, \Delta_k, A'_k)), \tag{2}$$

Here $\mathcal{U}$ applies edits and optional appearance replacement, and $\mathcal{P}$ normalizes and validates. Invalid or unchanged proposals return errors to the same reviewer while preserving $\mathcal{D}_k$. Whole-sequence

validation permits coordinated geometry and appearance changes within DSL and backend capabilities. E.g., replacing an object representation can require updating its appearance assignment in the same validated transaction. Physical-field changes trigger simulation; camera or appearance-only replays cached geometry from the current rollout, re-render controls, and repeat visual review.

Execution recovery and termination. Compilation, simulation, resource, and rendering errors prompt the author to correct the complete document. Authoring and review each allow up to three invalid-reply corrections. Invalid-response retries, asset-design attempts, frame requests, and scene-refinement rounds use separate budgets. Successful simulations are unlimited unless an explicit execution budget is set. The loop ends on acceptance or budget exhaustion. Accepted rollouts supply synthesis controls after the reviewer confirms event realization and visibility.

3.4 Simulation as a Video Prior

Aligned controls from a shared geometric sequence. We cache particle positions and solid surfaces at each output time. Liquid surfaces are extracted on CUDA from a smoothed particle-density grid using Marching Cubes (Lorensen & Cline, 1987), followed by weighted mesh and normal smoothing adapted from SplashSurf (Löschner et al., 2023); solids retain captured meshes. The same geometry supplies previews and controls. A shared camera and rasterizer produce depth D_t, view-space normals N_t, and entity labels L_t with consistent nearest-surface visibility. Contour maps B_t, supplied through VACE's scribble-control interface, encode silhouettes, occlusion boundaries, and depth/normal structure. Depth defaults to per-frame min–max normalization of valid non-plane metric depth, with nearer surfaces brighter; clip-wide normalization is optional (Appendix A.4). After acceptance, controls are validated and material masks derive from geometry and appearance assignments. Cached geometry permits reconstruction and rendering without repeating simulation.

Optional rollout-grounded captioning. Geometric controls are always derived from the accepted rollout. Optionally, the synthesis prompt is augmented with an event caption grounded in that rollout. Caption construction combines rollout evidence, the original request, and scene/appearance specifications through five stages: *objects* identifies first-frame participants and scene identities; *appearance* uses authored specifications and conservative category knowledge; *motion* describes sampled-frame dynamics and uncertainties; *context* incorporates scene-derived camera facts; and *compose* produces a narrative in event order. Checks enforce entity coverage and the text encoder's input budget (Appendix A.5).

From geometric controls to appearance. Pretrained VACE (Jiang et al., 2025) requires no additional training. Following Fluid Composer (Chen et al., 2025b), depth guides early denoising and contours guide later steps:

$$c_\ell = \begin{cases} D_{1:T}, & 0 \le \ell < K_D, \\ B_{1:T}, & K_D \le \ell < K, \end{cases} \qquad (K_D, K) = (4, 20), \tag{3}$$

Here T is the frame count, K the sampling-step count, and K_D the initial depth-conditioned steps. Depth establishes structure; contours emphasize detail.

Each appearance group, including background, defines a material branch. Its prompt x_i combines material text with a summarized rollout caption or, without caption, the original scene description. Geometry provides binary visibility masks M_i: opaque surfaces occlude, while transparent surfaces permit overlap. VAE temporal sampling and nearest-neighbor resizing yield latent masks $\widehat{M}_i$, which must cover the latent domain. Following Fluid Composer, complete branch updates combine as

$$w_i = \frac{\widehat{M}_i}{\sum_j \widehat{M}_j}, \qquad z_{\ell+1} = \sum_i w_i \odot \Phi_{\ell,i}(z_\ell; x_i, c_\ell), \tag{4}$$

where z_0 is Gaussian noise and $\Phi_{\ell,i}$ a guided material-branch update. Branches share z_ℓ but keep separate UniPC histories; overlapping binary masks receive equal weights. Mask overlap provides the compositional treatment of transparency. The final latent decodes to Y.

The default VACE-14B backend uses guidance 5 and timestep shift 16, generating 81 frames at 16 FPS and 832×480. Physical duration is independent of playback rate. Final-video evaluation

measures event fidelity after appearance synthesis. We evaluate two simulation-conditioned variants built from accepted rollouts: one retains the original scene description, and the other uses the optional rollout-grounded caption.

4 EXPERIMENTS

We compare final videos with Wan and analyze structured scene representation, iterative refinement, and optional caption conditioning. Simulation previews and refinement records provide evidence about scene construction, in addition to the final-video ratings.

4.1 EXPERIMENTAL SETUP

Physical task categories. We evaluate all 13 benchmark scenarios covering rigid-body and deformable contact, granular collapse, fluid flow, brittle fracture, constrained rigid motion, flexible-body motion, and multiphysics coupling. We use the four-configuration comparisons for these scenarios. Each benchmark scenario contributes equally. The complete benchmark is listed in App. C.1.

Baselines and input conditions. Wan 2.2 T2V and I2V use the *initial prompt*: the original request that initiates scene construction. I2V additionally receives the first frame extracted from our generated video for the corresponding scenario, providing its initial scene and appearance. Neither baseline receives rollout-derived captions or simulated control sequences. Our method starts from the same request and refines a physical scene whose rendered controls guide synthesis, with or without a rollout-derived caption. The comparison therefore evaluates complete configurations with different conditioning inputs.

Scope of baseline comparison. PhysCtrl, PhysGen3D, RealWonder, and PhysAgent (Wang et al., 2025; Chen et al., 2025a; Liu et al., 2026; Li et al., 2026) start from input images and use method-specific physical controls or event prompts, whereas our pipeline constructs the scene from text. These differences in input and control conditions preclude a matched quantitative comparison: pairing methods would require changing the input modality, the available physical controls, and the event specification simultaneously. Such a comparison would conflate method capabilities with unequal problem definitions. We focus here on Wan T2V and I2V as direct-generation baselines anchored to the same initial prompt (with I2V additionally receiving the first frame), and evaluate all configurations under the same four-criterion rubric.

Video backends. The official Wan 2.2 checkpoints used for the T2V and I2V baselines expose text and text-plus-single-image conditioning, respectively, rather than the VACE-style spatiotemporal control-video interface required by our synthesis stage. We therefore use pretrained Wan2.1-VACE-14B, whose documented depth and scribble-control pathways accommodate our temporally aligned depth and contour sequences (Wan Team et al., 2025; Jiang et al., 2025). Consequently, the comparison evaluates complete generation systems rather than isolating simulation conditioning under a shared video backbone.

Generation settings. Wan2.1-VACE-14B uses 20 UniPC steps, timestep shift 16, seed 2026, and depth conditioning for the first 20% of the schedule, followed by contours. In the reported configurations, Ours w/o caption uses the input prompt directly, whereas Ours w/ caption additionally includes a rollout-grounded caption. Evaluated clips have 80 or 81 frames at 16 FPS and 832×480. Representation and refinement analyses use separate simulation previews whose presentation rendering preserves cached motion and geometry.

Evaluation. We request `gpt-5.6-sol` (regular route, high reasoning effort) to assign integer physical-fidelity ratings from 1 to 5 for object preservation (OP), contact and event-direction correctness (CD), support and gravity consistency (SG), and physical-response plausibility (PR). All original frames are supplied chronologically as adjacent full-frame pairs with timestamps and no method labels. The evaluator receives the event description and uses the first video frame as an initial-scene proxy; no independent conditioning image is supplied. For each scenario, the same target event description is used for all four configurations. Each final video receives three independent

evaluator calls. For call r, we average the four raw ratings, $S_v^{(r)} = \frac{1}{4}\sum_d s_{v,d}^{(r)}$; we then average the three calls per video and weight all 13 scenarios equally. Reported standard deviations are the sample standard deviation across the three call-level method means, after each call is averaged over the same scenarios. They describe repeat-evaluation variation rather than variation in scenario difficulty. App. C.2 details the input protocol and scoring rubric.

4.2 Comparison with Wan

Quantitative comparison. Table 1 reports mean ratings of 3.69 without caption and 3.56 with caption, versus 3.14 for Wan T2V and 3.09 for Wan I2V. Both of our configurations outperform both direct-generation baselines in the aggregate. Relative to T2V, the largest gains are in contact and event direction, followed by physical response; support and gravity remain closer. The strongest gains therefore concern realizing the requested physical event and its response.

Table 1: **Quantitative comparison.** We compare final-video ratings across all 13 benchmark scenarios. Wan baselines use the initial prompt; I2V also receives the first frame of our generated video. Every entry reports the three-call mean and the sample standard deviation across the three call-level method means. **Bold** and underlined denote the best and second-best results, respectively.

Method	OP ↑	CD ↑	SG ↑	PR ↑	Mean ↑
Wan 2.2 T2V	4.13±0.32	2.69±0.13	3.44±0.04	2.31±0.08	3.14±0.11
Wan 2.2 I2V	3.67±0.12	3.21±0.04	3.23±0.08	2.26±0.04	3.09±0.02
Ours w/o caption	**4.15±0.13**	3.97±0.16	**3.72±0.12**	**2.90±0.09**	**3.69±0.07**
Ours w/ caption	4.08±0.08	**4.03±0.22**	3.54±0.08	2.59±0.16	3.56±0.10

Qualitative comparison. Fig. 2 compares rigid contact and brittle fracture. In the two-ball collision example, T2V adds balls and shows discontinuous or weakly caused motion. Ours preserves two balls and post-contact separation, increasing the score from 2.92 to 4.42. For the vase, T2V shows an airborne initial state and suspended fragments. Ours more clearly realizes falling and breaking, increasing the score from 2.08 to 3.08.

4.3 Ablation Studies and Analysis

We ablate three design choices: the structured scene representation, iterative review and refinement, and optional rollout-grounded captioning. The first two affect scene-authoring reliability, whereas captioning affects only final video synthesis. For the authoring ablations, we report accepted-within-budget success separately from preview quality: a run may produce a complete, rateable preview without satisfying the acceptance criterion within its budget. Preview scores use the same four-dimensional rubric as above but come from a separate frozen evaluation of the simulation previews.

Table 2: **Authoring ablations.** (a) Direct Genesis v.s. our structured representation, compared on all 13 scenarios; preview quality uses the last complete preview from each arm. (b) Before and after iterative refinement on the same runs; Loop quality uses the nine scenarios with rateable previews at both endpoints. Three initial runs have no complete video, and the initial sand-rabbit video is excluded because its below-table camera leaves the motion occluded; these cases are not assigned zero scores.

(a)	Accepted	OP	CD	SG	PR	Mean
Direct Genesis	7/13	4.23	3.23	4.00	2.69	3.54
Structured DSL	**13/13**	**4.92**	**4.15**	4.00	**3.69**	**4.19**

(b)	Accepted	Complete	Quality
Loop off	1/13	10/13	3.50
Loop on	**13/13**	**13/13**	**4.25**

Structured scene representation ablation. We compare our structured representation with direct Genesis code generation by the same model. Both arms use the same task descriptions, execution budget, online execution and visual feedback, and acceptance criterion. The structured representation raises accepted-within-budget success from 7/13 to 13/13 and the last-complete-preview mean from 3.54 to 4.19. It improves nine scenarios, ties two, and declines on two; the largest aggregate gains are in physical response and contact/event direction, while support and gravity are unchanged. A complete preview can receive a quality score even when the run fails reviewer acceptance. App. D.1 reports per-scenario outcomes and a pillow comparison.

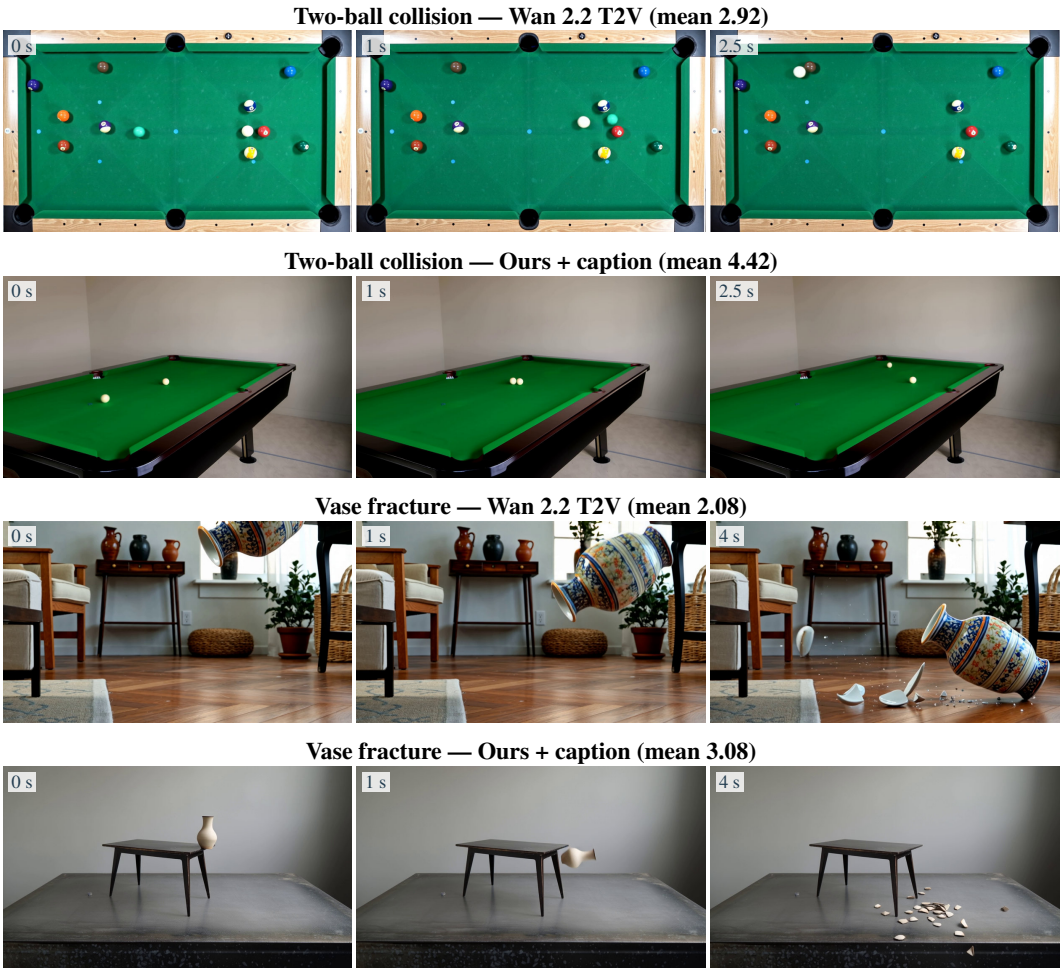


Figure 2: **Qualitative comparison.** For each scenario, the upper row is Wan 2.2 T2V and the lower row is ours with caption-conditioned synthesis; frames advance left to right. Scores average the four criteria and three evaluator calls over the full video.

Iterative review and refinement. Loop off stops after the initial rollout, before reviewer-driven revision; loop on uses the final accepted rollout from the same run. Only one of 13 initial rollouts is accepted, whereas refinement brings all 13 to acceptance. Among the nine scenarios with rateable previews at both endpoints, mean quality increases from 3.50 to 4.25 ($+0.75$): six improve, two tie, and one declines. Initial failures remain part of the 13-scenario success analysis but are not mapped to artificial zeros on the 1–5 quality scale. The initial sand-rabbit video is complete, but its camera lies below the tabletop, which occludes the rabbit's collapse and the resulting sand pile; consequently, the four physical criteria cannot be scored from that video. Across the 13 scenarios, the median first-acceptance round is five and the maximum is nine. Fig. 6 illustrates the basketball bank shot: the initial rollout contacts the backboard but misses the rim, whereas the round-6 rollout shows the accepted backboard-to-rim passage. Intermediate revisions address the launch state, hoop asset/orientation, and camera. App. D.2 reports per-scenario endpoints and first-acceptance rounds; App. D.3 details the intermediate diagnoses and review actions.

Impact of optional rollout-derived captioning. Geometric controls guide motion and structure but leave appearance synthesis partly to the video model's learned priors. The model may therefore introduce content absent from the simulated scene, affecting object identity and event fidelity. An optional rollout-grounded caption explicitly describes the participants and their interaction, providing a semantic constraint to reduce such additions. Fig. 4 illustrates this role in billiards. Without a caption, synthesis adds a small blue-green, ball-like detail absent from the simulated scene. With a caption, this region appears as a small surface mark rather than an additional ball-like object, while

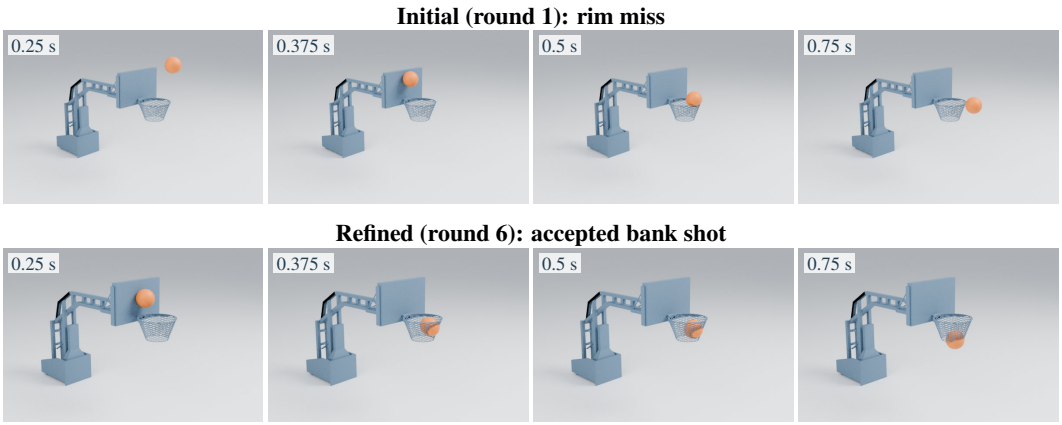


Figure 3: **Simulation-in-the-loop refinement of a basketball bank shot.** The initial rollout (top) contacts the backboard but misses the rim; the accepted rollout (round 6, bottom) realizes the request. Labels indicate playback time. Intermediate diagnoses and review actions are reported in App. D.3.



Figure 4: **Optional caption correction in billiards.** Matched frames at 0.5 s, without (left) and with (right) a rollout-grounded caption. Circles mark the blue-green ball-like detail and the corresponding surface mark. Scores are three-judgment means over the full videos.

preserving the intended two-ball interaction. Object preservation improves from 4.00 to 5.00, and the full-video mean rises from 4.08 to 4.42. Captioning remains optional: across scenarios, its mean is 3.56 versus 3.69 without it, supporting its use for semantic correction.

5 Discussion & Conclusion

We presented an agentic framework for text-to-video synthesis that constructs, validates, and simulates an explicit physical world before generating its appearance. A structured scene representation makes interaction geometry, material properties, and physical roles editable, while a vision-language controller uses simulated rollouts to assess the requested event and revise the scene through validated edits and renewed execution. The accepted geometry then supplies aligned depth, contour, and material controls for pretrained video synthesis without additional model training. Experiments across diverse physical scenarios show higher final-video physical fidelity than ablated variants and direct text-guided video generation baselines, even when the latter receive additional first-frame conditioning. These findings support constructing and refining the physical world underlying an event as a basis for faithful video generation.

The framework remains limited by the materials and interactions supported by its simulators, the reliability of visual rollout assessment, and the cost of repeated simulation. Acceptance reflects the controller's judgment of the observed event, and generative synthesis can still alter motion despite geometric conditioning. Extending the supported physical interactions, improving refinement efficiency, and checking event preservation in generated videos are natural directions for future work.

AI USE STATEMENT

Generative AI is an integral component of our method: LLM/VLM agents construct and refine executable physical scenes, inspect simulated rollouts, and optionally produce rollout-grounded captions; a pretrained video model synthesizes the final appearance. We also use a vision-language model for fixed-rubric evaluation of the generated videos. These uses are explicitly described in the main paper and supplementary material.

Beyond the proposed system, we **used generative AI tools to** assist with literature search, reference formatting, manuscript editing and translation, creating or modifying scientific figures, intermediate experimental analysis, and designing the webpage for the supplemental material. **We have reviewed all AI-assisted work.** AI-assisted suggestions, code, and result summaries were reviewed and finalized by the authors; AI created or modified figures are recreated, modified and verified by authors. Writing suggestions on reported measurements, interpretations, and scientific conclusions were verified by the authors. AI-suggested related work was checked against the original sources. LLM-generated code was verified and tested for correctness by three authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

We have not used generative AI tools for these tasks: to help develop theoretical models or conceptual frameworks, or to design or provide feedback on research methodology or experiments. **The following subtasks are not applicable to this work:** to generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, assist in writing proofs, or clean or reformat datasets.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document the experimental configurations, video-generation settings, and evaluation procedure in Sec. 4.1. Appendix A specifies the scene representation, validation rules, physics backends, resource bounds, and geometry-to-video controls, while Appendix B summarizes the agent decision rules and prompts. The complete prompt templates and schemas are provided in the supplementary materials. Appendix C details the benchmark, evaluation protocol, and complete per-scenario scores; the supplementary materials additionally include evaluation prompts, prompt-version metadata, scored-video hashes, and criterion-level rationales to support verification of the reported results. We will release the source code as open source upon acceptance of the paper.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A Space-Time Diffusion Model for Video Generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11. Association for Computing Machinery, 2024. ISBN 9798400711312. doi: 10.1145/3680528.3687614.

S. G. Bardenhagen, J. U. Brackbill, and D. Sulsky. The material-point method for granular materials. *Computer Methods in Applied Mechanics and Engineering*, 187(3–4):529–541, 2000. doi: 10.1016/S0045-7825(99)00338-2.

R. L. Bishop. There is more than one way to frame a curve. *The American Mathematical Monthly*, 82(3):246–251, 1975. doi: 10.1080/00029890.1975.11993807.

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*, 2023. doi: 10.48550/arXiv.2311.15127.

E. Catmull and R. Rom. A class of local interpolating splines. In R. E. Barnhill and R. F. Riesenfeld (eds.), *Computer Aided Geometric Design*, pp. 317–326. Academic Press, 1974. doi: 10.1016/B978-0-12-079050-0.50020-5.

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. PhysGen3D: Crafting a Miniature Interactive World from a Single Image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6178–6189, 2025a. doi: 10.1109/CVPR52734.2025.00579.

Duowen Chen, Zhiqiang Lao, Yu Guo, and Heather Yu. Fluid composer: Fluid detail composition and rendering using video diffusion models. *Computer Graphics Forum*, pp. e70300, 2025b. doi: 10.1111/cgf.70300.

P. A. Cundall and O. D. L. Strack. A discrete numerical model for granular assemblies. *Géotechnique*, 29(1):47–65, 1979. doi: 10.1680/geot.1979.29.1.47.

Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond. Computer software, 2024.

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*, volume 38, pp. 103174–103201, 2025. doi: 10.52202/085713-3448.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, Charles Herrmann, 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*, pp. 20077–20087, 2026.

Yunqi Gu, Ian Huang, Jihyeon Je, Guandao Yang, and Leonidas Guibas. BlenderGym: Benchmarking Foundational Model Systems for Graphics Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18574–18583, 2025a. doi: 10.1109/CVPR52734.2025.01731.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. In *Proceedings of the SIGGRAPH Conference Papers*, SIGGRAPH Conference Papers '25. Association for Computing Machinery, 2025b. ISBN 9798400715402. doi: 10.1145/3721238.3730607.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *The Twelfth International Conference on Learning Representations*, volume 2024, pp. 34630–34648, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High Definition Video Generation with Diffusion Models. *arXiv preprint arXiv:2210.02303*, 2022a. doi: 10.48550/arXiv.2210.02303.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022b. doi: 10.52202/068431-0628.

Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A. Ross, Cordelia Schmid, and Alireza Fathi. SceneCraft: An LLM agent for synthesizing 3D scenes as blender code. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19252–19282. PMLR, 2024.

Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The Material Point Method for Simulating Continuum Materials. In *ACM SIGGRAPH Courses*, SIGGRAPH '16. Association for Computing Machinery, 2016. ISBN 9781450342896. doi: 10.1145/2897826.2927348.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025. doi: 10.1109/ICCV51701.2025.01597.

Qirui Li, Jinkun Hao, Yibo Li, Ran Yi, Paul L. Rosin, and Yu-Kun Lai. PhysAgent: Reflective agentic physics control for physically plausible video generation. *arXiv preprint arXiv:2607.16355*, 2026. doi: 10.48550/arXiv.2607.16355.

Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *Computer Vision – ECCV 2024*, volume 15140 of *Lecture Notes in Computer Science*, pp. 360–378. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73007-8_21.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. RealWonder: Real-Time Physical Action-Conditioned Video Generation. *arXiv preprint arXiv:2603.05449*, 2026. doi: 10.48550/arXiv.2603.05449.

W. E. Lorensen and H. E. Cline. Marching Cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987. doi: 10.1145/37402.37422.

Sining Lu, Guan Chen, Nam Anh Dinh, Itai Lang, Ari Holtzman, and Rana Hanocka. LL3M: Large Language 3D Modelers. *arXiv preprint arXiv:2508.08228*, 2025. doi: 10.48550/arXiv.2508.08228.

Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. GPT4Motion: Scripting Physical Motions in Text-to-Video Generation via Blender-Oriented GPT Planning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1430–1440, 2024. doi: 10.1109/CVPRW63382.2024.00150.

Fabian Löschner, Timna Böttcher, Stefan Rhys Jeske, and Jan Bender. Weighted Laplacian Smoothing for Surface Reconstruction of Particle-based Fluids. In *Vision, Modeling, and Visualization*. The Eurographics Association, 2023. ISBN 978-3-03868-232-5. doi: 10.2312/vmv.20231245.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 2025.

Matthias Müller, David Charypar, and Markus Gross. Particle-Based Fluid Simulation for Interactive Applications. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '03, pp. 154–159. Eurographics Association, 2003. ISBN 1581136595.

Saurabh Pathak, Elahe Arani, Mykola Pechenizkiy, and Bahram Zonooz. PhysVid: Physics Aware Local Conditioning for Generative Video Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41847–41858, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377.

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. In *The Eleventh International Conference on Learning Representations*, 2023.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. doi: 10.48550/arXiv.2503.20314.

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*, volume 38, pp. 167907–167932, 2025. doi: 10.52202/085713-5598.

Rui-Huan Wang, Si-Tong Wei, Jia-Qi He, Heng-Yi Wei, Baoquan Chen, and Peng-Shuai Wang. aDSL: Agentic 3d creation via joint agent-program design. *arXiv preprint arXiv:2608.17975*, 2026. doi: 10.48550/arXiv.2608.17975.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional Video Synthesis with Motion Controllability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0334.

Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. PhysAnimator: Physics-Guided Generative Cartoon Animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10793–10804, 2025. doi: 10.1109/CVPR52734.2025.01008.

Jingrui Xing, Liangwang Ruan, Bin Wang, Bo Zhu, and Baoquan Chen. Position-Based Surface Tension Flow. *ACM Transactions on Graphics*, 41(6), 2022. doi: 10.1145/3550454.3555476.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.

Shaofeng Yin, Jiaxin Ge, Zora Zhiruo Wang, Chenyang Wang, Xiuyu Li, Michael J. Black, Trevor Darrell, Angjoo Kanazawa, and Haiwen Feng. Vision-as-Inverse-Graphics Agent via Interleaved Multimodal Reasoning. In *Computer Vision – ECCV 2026*, pp. 557–575. Springer Nature Switzerland, 2026. ISBN 978-3-032-37464-6. doi: 10.1007/978-3-032-37464-6_30.

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 *The Fourteenth International Conference on Learning Representations*, 2026.

Hongxin Zhang, Chunru Lin, Junyan Li, Zhou Xian, Tsun-Hsuan Wang, and Chuang Gan. GS-Agent: Creating 4D physical worlds with generative simulation. *arXiv preprint arXiv:2607.21522*, 2026. doi: 10.48550/arXiv.2607.21522.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3813–3824, 2023. doi: 10.1109/ICCV51070.2023.00355.

T. Y. Zhang and C. Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984. doi: 10.1145/357994.358023.

A Method Specification and Implementation Scope

This appendix specifies the representation, simulation scope, validation checks, rendering controls, and rollout-caption construction used by the system. The organization follows the execution and evidence path from a scene contract to an accepted rollout and its synthesized video; numerical limits are the defaults shipped with the system.

A.1 Scene Representation and Validation

Asset contracts and scene instances. Asset contracts describe geometry and part roles, permitted scales and materials, available support surfaces, and source anchors. Their fields depend on asset type; no single asset is assumed to provide every capability. Fixed rigid assets may have independent visual and collision representations. Articulated assets additionally specify link hierarchy, masses, inertias, and joint constraints. Elastic assets require valid volumetric geometry whose surface can follow deformation, together with effective Young's modulus E, Poisson's ratio ν, and density ρ.

The scene instance selects pose, scale, physical role, initial conditions, and motion controls within that contract. Validation and provenance records retain asset sources and file hashes, the simulator and version, probe mode and tested scale, and stated assumptions, usage, and limitations. These records establish the scope of compatibility checks; they do not introduce physical degrees of freedom. Registration compatibility does not establish interaction success; success is assessed from the complete rollout.

Articulated mechanisms. An articulated asset is specified as rigid parts and joints. Deterministic checks enforce valid dimensions, masses, and a rooted tree structure before compilation, and analytical formulas determine each part's inertia. Neutral multiview renders are reviewed for proportions, recognizable shape, completeness, and relative part sizes; this review concerns geometry rather than dynamics. A separate simulator probe tests loading, scene construction, prescribed initial joint states, and a short rollout with finite joint states. Validation, rendering, visual-review, and probe failures return structured feedback and the previous design to the generator. All failure types share a budget of five design attempts, and only a design that passes the checks is registered.

Generated mechanisms have a fixed base and a tree of rigid parts linked by joints, with at most one connect constraint closing a loop. These checks assess geometric suitability and limited simulator compatibility, not long-term stability, general contact behavior, or successful realization of the requested event in a complete scene. Ropes are simulated as segmented articulated chains. For rendering, the per-frame endpoints are joined into a polyline, resampled with a Catmull–Rom spline (Catmull & Rom, 1974), and swept into a tube by a rotation-minimizing frame (Bishop, 1975), closed with hemispherical end caps. This is a discrete chain approximation, not a continuum cable solver.

A.2 Physics Backends and Supported Scope

The system combines native Genesis rigid-body simulation with project-specific extensions. The following scope statements distinguish cited background methods from project-specific approximations.

Liquid and separable MPM contact. The project-local PBSTF solver (Xing et al., 2022) runs alongside unmodified Genesis (Genesis Authors, 2024). Validation enforces a single liquid family per scene (MPM (Jiang et al., 2016), SPH (Müller et al., 2003), and PBSTF liquids are not mixed)

and restricts this adapter to one-way interaction with fixed rigid colliders; scenes containing dynamic rigid bodies or articulated URDF assets are rejected with an explicit error rather than silently adapted.

Standard MPM (Jiang et al., 2016) and multi-material contact (Bardenhagen et al., 2000) provide the background; the following behavior is project-specific. Independent MPM elastic bodies accumulate per-body node mass and momentum. Contact applies equal-and-opposite unilateral compressive impulses with Coulomb-limited tangential friction. Proximity tests are conservative grid-scale bounds derived from particle supports, simultaneous multi-body contacts are resolved with four rounds of local projection, and coupling to rigid bodies runs once per outer step through the native solver. The extension is forward-only, and an implementation fingerprint is recorded with each run.

Geometry-dependent fluid resolution. For MPM liquids, the author specifies accuracy targets (N_W, N_F). A separate step identifies interaction width W and the smallest relevant initial fluid feature F, such as a source width or liquid-layer thickness, from resolved geometry and records the geometric justification. We derive

$$n_g = \left\lceil \max\left(\frac{N_W}{W}, \frac{N_F}{F}\right) \right\rceil, \qquad \Delta x = n_g^{-1}, \qquad (p, h_{\mathrm{rec}}, \ell_{\mathrm{mc}}) = (\alpha, \beta, \gamma)\Delta x, \tag{5}$$

where n_g is grid density, Δx grid spacing, p particle spacing, h_{rec} reconstruction smoothing length, and ℓ_{mc} reconstruction grid spacing. Defaults are $(N_W, N_F) = (32, 4)$ and $(\alpha, \beta, \gamma) = (0.4, 0.65, 0.25)$. Scales are recomputed before rollouts; geometric changes repeat feature identification, while camera-only edits reuse established scales. This rule targets initial features, not future splash-sheet thickness. Conservative sampling bounds, emission duration, speed, and particle spacing determine emitter storage without changing authored flow rates.

Granular material and fracture. DEM sand uses a discrete-element representation (Cundall & Strack, 1979) and is volume-sampled into grains inside a dedicated domain box, with the initial grain count capped at 5×10^5. The domain must not cut into sand shapes, separate sand shapes must be spaced by at least one grain diameter, and dynamic rigid coupling is rejected.

Fracturable objects carry deterministic voxel–Voronoi pre-fractures with adjacent fragments sharing coplanar, non-overlapping faces. Three release modes are supported: *deferred* (default), *welded*, and *unbonded*; in deferred mode the intact mesh falls and fragments replace it upon first contact with an exterior support, welded shards release on first novel contact, and unbonded shards start loose. Release can additionally be gated on an accumulated contact-impulse threshold; resting support contacts close the impulse window without firing, so object weight alone never triggers fracture.

Contact parameters and project-specific limits. Authored `contact_dampratio` values are floored for stability at 0.5 under a hard contact time constant ($\leq 2\times$ the solver substep) and at 0.2 on the default soft contact; the floors come from pendulum and gap-chain parameter sweeps. Pairwise friction is the maximum of the two surfaces' values with an effective floor of 0.01. These settings and backend restrictions define the supported scope; they are not claims of a general multi-physics solver.

A.3 Closed-Loop Checks and Resource Bounds

Review gates and measurements. Acceptance combines deterministic validation, numerical checks, and visual review of the complete rollout. The numerical diagnostics include a floor-crossing check that flags a dynamic rigid body whose rendered visual mesh lies entirely below a fixed horizontal floor by more than 1 mm, measured from cached world-space mesh vertices. Additional measurements supplied to the reviewer include world-coordinate geometry, soft-body approach evidence, and articulation degrees of freedom read from the joint-state log. Together these checks support the rollout review; asset registration or a short probe alone does not determine whether the requested interaction is realized.

Execution budgets. Scenes are subject to fixed resource budgets: at most 2.4×10^7 grid cells, 2×10^6 initial particles (5×10^5 for DEM grains), 10^6 particles per emitter, and 3×10^6 particles in total; at most 2×10^5 steps and 2×10^6 substeps; and memory budgets of 12 GiB for elastic support and 4 GiB for separable contact.

A.4 GEOMETRY-TO-VIDEO CONTROLS

Liquid surfaces are reconstructed by a CUDA density-field pipeline with Marching Cubes (Lorensen & Cline, 1987) by default, with SplashSurf (Löschner et al., 2023) available as an explicit reference backend. Depth controls use a VACE-style conditioning interface (Jiang et al., 2025). By default, each frame's valid non-plane metric depths set the min–max range; depth is mapped linearly to near-white/far-black intensities, missing pixels are black, and no reciprocal-depth transform is used. Planes use the fitted range but do not determine it; if no non-plane foreground is visible, all valid geometry is used. The optional clip-wide mode fits a shared range to the non-plane foreground's 1% and 99.5% depth percentiles, with the same fallback.

Contour maps are supplied through VACE's scribble-control interface and extracted from geometry alone: silhouettes and occlusion boundaries from relative depth jumps, and orientation structure from multi-scale normal contrast with radial-minima selection, thinned to skeleton lines (Zhang & Suen, 1984). Material masks are exported from captured or reconstructed geometry with appearance assignments, reusing the same per-group base color as neutral diffuse preview shading.

A.5 OPTIONAL ROLLOUT-GROUNDED CAPTION CONSTRUCTION

For caption-conditioned synthesis, the vision-based *objects* stage receives the original request, accepted scene JSON, video metadata, and first rendered frame. It inventories visible or soon-visible entities with identifiers, natural names, and coverage nouns while ignoring neutral colors and materials. The vision-based *appearance* stage receives this inventory, request, scene JSON, metadata, appearance JSON, and first frame; the appearance specification is authoritative and conservative category knowledge fills missing cues.

The vision-based *motion* stage receives the object inventory, request, scene JSON, metadata, and approximately twelve evenly sampled frames. It describes observed dynamics and interactions, treats liquids as continuous flows, and records uncertainty without claiming unsampled events. The text-only *context* stage combines these outputs with a camera sentence derived programmatically from the scene camera, inferring venue, background, camera wording, and lighting conservatively.

The text-only *compose* stage receives the request, prior outputs, camera sentence, and coverage nouns. It produces a paragraph in chronological order, covers every visible entity, uses appearance only from appearance decisions, and excludes production or meta-language. Intermediate outputs and quality-check traces are recorded; final checks cover entity coverage, single-paragraph form and text-encoder budget. Before VACE conditioning in caption-conditioned runs, a distinct summarization pass condenses the caption while retaining subject identity, essential material, causal interaction, and motion; this preparation is separate from the five caption stages.

B AGENT DECISION RULES AND PROMPT SUMMARY

This appendix summarizes the decision rules used by five functional modules; it is not a verbatim listing of the API prompts. At runtime, template fields are filled with the request, scene state, available evidence, and images required by each stage.

B.1 WORLD CONSTRUCTION AND ASSET SUITABILITY

The construction stage turns the request into an executable scene using SI units and z-up coordinates. It preserves requested identities, counts, materials, constraints, initiating motion, decisive interaction, and visible aftermath. Simple objects may use primitives; distinctive geometry and mechanisms use asset tools. Independent motion requires independent bodies, required joints cannot be replaced by a static mesh, and explicit compression or recovery requires a deformable representation. Passive motion must arise from simulation rather than repeated velocity resets or disabled collisions. Camera placement must show cause, interaction, and response.

Geometry and mechanics are returned separately from appearance. Each object or emitter is assigned to a material group with an appearance description, together with a background group. Appearance text cannot supply missing geometry or physical behavior.

Table 3: **Functional modules and decision rules.** The table summarizes the inputs, outputs, and core decisions of the five functional modules.

Stage	Inputs	Output	Core decision
World construction	request, scene interface, asset capabilities	executable scene and appearance groups	separate geometry and mechanics from appearance
Asset acquisition	asset request, catalog results, measured views, interaction requirements	search, inspect, register, or fail action	verify static functional suitability before registration
Asset review	request, asset views, measurements, interfaces	per-object match and missing requirements	separate intrinsic geometry from dynamic outcome
Rollout review	request, scene state, frames, measurements, feedback	request frames, revise, repair, accept, or request review	act only on observed evidence
Caption construction	request, inventory, grounded appearance, motion, context, camera	one paragraph	describe the accepted event causally without unsupported details

Asset acquisition and review establish whether an asset can support the requested interaction; a successful download or recognizable name is not evidence of suitability. The agent checks functional dimensions, openings, contact surfaces, joints, orientation, material capabilities, and collision requirements in world coordinates. It records supported scales, part roles, support surfaces, source anchors, material assessment, and limitations before registration. A candidate may be suitable while its eventual dynamic trajectory remains unverified. Placement defects are handled by valid transforms; intrinsic geometry or interface defects require replacement or repair.

B.2 Rollout Review and Revision

The rollout reviewer receives the original request, current scene and appearance, metadata, asset interfaces, executed measurements, prior feedback, and visual evidence. It first requests preview frames when none are attached, then requests exact or adjacent frames when contact, occlusion, or fast motion is ambiguous. It never accepts or revises an outcome without seeing the relevant frames.

The reviewer checks recognizable participants, initiating cause, decisive interaction, and visible aftermath. Before tuning dynamics, it checks placement, orientation, functional faces, clearance, and the complete contact path. Scene edits address placement, material, timing, or camera problems; asset repair addresses intrinsic geometry or interface defects; camera-only changes may reuse cached physics. Each revision states the observed failure, its evidence and cause, the proposed correction, and the expected visible change. The reviewer must not infer an occluded interaction from the final arrangement, hide a failure by cropping, or treat successful loading as task success.

The action boundary is explicit: `request_frames` asks for indexed evidence; `revise` applies validated scene edits; `repair_assets` identifies object-specific intrinsic defects; `accept` records observed evidence that the event is realized and readable; and `needs_review` reports an unsupported capability or missing input. Ordinary repairable failures lead to revision rather than stopping.

B.3 Optional Caption Construction

When this optional component is enabled, caption construction first produces an object inventory, grounded appearance decisions, a chronological motion narrative, conservative context, and an authoritative camera description. The inventory and appearance stages use the first frame; motion uses sampled chronological frames. The final stage combines these structured outputs with the request to write a paragraph for the accepted event.

The caption follows causal order: initiation, interaction, and visible consequences. It keeps entity names, quantities, arrangements, appearance, and informative final states consistent through motion, deformation, and occlusion. It uses only grounded appearance decisions, omits invisible helpers and production steps, and does not infer material or texture from neutral shading. It may describe

supported setting, camera, and simple lighting naturally, but must not add objects, causes, events, or unsupported details. The output does not mention simulation, rendering, meshes, or hidden asset construction; subsequent text preparation handles the complete video-conditioning budget.

C Evaluation Protocol and Complete Results

This appendix defines the benchmark and evaluation protocol before reporting complete per-scenario results. The quantitative evaluation concerns final appearance-synthesized videos, not simulation previews. It therefore measures complete generation configurations rather than simulation accuracy in isolation.

C.1 Benchmark and Compared Configurations

The benchmark contains all 13 scenarios used in the quantitative comparison. We use the following fixed, result-independent order throughout this appendix.

Table 4: **Benchmark scenarios.** The complete benchmark contains 13 scenarios grouped by physical phenomenon.

Physical phenomenon	Scenarios
Rigid-body contact	Dominoes; two-ball collision; bowling
Constrained/flexible motion	Newton's cradle; hanging lanterns
Deformable contact	Basketball bank shot; pillows
Granular collapse	Sand rabbit
Fluid and multiphysics	Champagne tower; water wheel; ball and mattress
Brittle fracture	Vase fracture; ball impacts glass tower

The four configurations are summarized below. Wan 2.2 T2V and I2V are direct-generation baselines. Our two configurations share the accepted simulation controls and synthesis settings; they differ in whether the text condition uses the original scene description or a rollout-derived caption.

Table 5: **Compared configurations.** The table summarizes the inputs and conditioning used by the four evaluated configurations.

Configuration	Backend	Inputs and conditioning
Wan 2.2 T2V	Wan 2.2 T2V	Initial text request
Wan 2.2 I2V	Wan 2.2 I2V	Initial text request and the first frame of the corresponding generated video
Ours w/o caption	Wan2.1-VACE	Accepted depth/contour controls and the original scene description
Ours w/ caption	Wan2.1-VACE	Accepted depth/contour controls and a rollout-derived event description

C.2 Final-Video Evaluation Protocol

We evaluate 52 videos: 13 scenarios under four configurations. Every evaluation requests `gpt-5.6-sol` through the AIHubMix Responses API using the same regular route and high reasoning effort. Each video is evaluated in three independent evaluator calls, without method labels or competing scores. All 80 or 81 frames are submitted chronologically at native 832×480 resolution, with two consecutive frames stacked vertically per input image, frame indices and timestamps at 16 FPS, and high image detail. The event description accompanies the frames, and the first frame serves as the initial-scene reference.

The evaluator assigns one integer from 1 to 5 for each of four criteria: object preservation (OP), contact and event-direction correctness (CD), support and gravity consistency (SG), and physical-response plausibility (PR). It is instructed to judge observable event alignment and physical plausibility over the full sequence, without rewarding visual quality when the event is wrong or inferring hidden simulator state. For video v and evaluator call r, the call score is $S_{v,r} = \frac{1}{4}\sum_d s_{v,r,d}$. The

reported video score is the mean of the three call scores, and method means weight all 13 scenarios equally. The reported standard deviation is the sample standard deviation ($n - 1$) of the three round-level method means.

All reported aggregates use the same four criterion definitions and rating anchors. Auxiliary logged fields are excluded from every table, figure, and aggregate. Complete per-scenario three-call means are reported below.

C.3 Aggregate and Per-Scenario Results

Table 1 in the main paper reports aggregate means. Ours with caption scores higher than T2V in 9 scenarios, ties in none, and scores lower in 4; against I2V the counts are 10, 1, and 2. Relative to ours without caption, the caption configuration wins in 5 scenarios, ties in none, and loses in 8. Its aggregate mean changes from 3.69 to 3.56; CD increases slightly, while OP, SG, and PR decrease.

Table 6: **Per-scenario results.** Three-call mean final-video ratings are reported for all 13 benchmark scenarios.

Scenario	Configuration	OP	CD	SG	PR	Mean
Dominoes	T2V	4.00	5.00	4.67	4.00	4.42
	I2V	5.00	3.67	3.67	2.67	3.75
	Ours w/o caption	5.00	5.00	5.00	4.00	4.75
	Ours w/ caption	4.00	4.67	4.67	3.33	4.17
Two-ball collision	T2V	2.67	4.00	3.00	2.00	2.92
	I2V	5.00	1.67	5.00	1.33	3.25
	Ours w/o caption	4.00	5.00	3.67	3.67	4.08
	Ours w/ caption	5.00	5.00	4.00	3.67	4.42
Bowling	T2V	3.67	2.33	2.00	2.00	2.50
	I2V	1.00	2.00	1.33	1.00	1.33
	Ours w/o caption	2.00	3.00	2.33	2.00	2.33
	Ours w/ caption	2.00	3.00	2.00	2.00	2.25
Newton's cradle (2 balls)	T2V	4.00	2.67	4.00	2.00	3.17
	I2V	3.00	4.00	3.67	3.00	3.42
	Ours w/o caption	5.00	3.00	5.00	2.33	3.83
	Ours w/ caption	5.00	3.00	4.67	2.33	3.75
Hanging lanterns	T2V	5.00	3.00	5.00	3.67	4.17
	I2V	3.33	4.00	4.00	3.33	3.67
	Ours w/o caption	5.00	4.33	5.00	4.00	4.58
	Ours w/ caption	3.00	3.67	4.00	3.00	3.42
Basketball bank shot	T2V	4.00	2.67	2.33	2.00	2.75
	I2V	4.00	3.67	3.33	3.00	3.50
	Ours w/o caption	4.67	5.00	4.33	3.67	4.42
	Ours w/ caption	4.67	4.67	4.33	3.33	4.25
Pillows	T2V	4.33	1.00	2.33	2.00	2.42
	I2V	2.33	4.00	2.67	2.00	2.75
	Ours w/o caption	4.67	3.00	2.00	3.00	3.17
	Ours w/ caption	4.67	3.00	2.00	2.00	2.92
Sand rabbit	T2V	5.00	3.00	4.33	2.00	3.58
	I2V	4.67	3.00	3.33	2.00	3.25
	Ours w/o caption	5.00	5.00	4.67	3.00	4.42
	Ours w/ caption	5.00	5.00	5.00	3.33	4.58
Champagne tower	T2V	4.00	2.67	4.00	3.00	3.42
	I2V	5.00	2.33	3.00	2.00	3.08
	Ours w/o caption	4.33	2.33	3.00	2.00	2.92
	Ours w/ caption	5.00	3.00	2.67	2.00	3.17
Water wheel	T2V	5.00	2.00	5.00	2.00	3.50
	I2V	3.00	2.33	4.33	2.00	2.92
	Ours w/o caption	3.33	2.67	4.00	2.33	3.08
	Ours w/ caption	3.33	4.00	4.33	2.67	3.58

Continued on next page

Table 6 (continued)

Scenario	Configuration	OP	CD	SG	PR	Mean
Ball and mattress	T2V	4.33	2.33	4.67	2.67	3.50
	I2V	4.67	4.67	4.67	4.00	4.50
	Ours w/o caption	5.00	5.00	4.33	4.00	4.58
	Ours w/ caption	5.00	5.00	3.00	2.33	3.83
Vase fracture	T2V	4.00	2.33	1.00	1.00	2.08
	I2V	3.33	2.33	1.00	1.00	1.92
	Ours w/o caption	4.00	4.00	3.67	2.33	3.50
	Ours w/ caption	3.33	3.33	3.67	2.00	3.08
Ball impacts glass tower	T2V	3.67	2.00	2.33	1.67	2.42
	I2V	3.33	4.00	2.00	2.00	2.83
	Ours w/o caption	2.00	4.33	1.33	1.33	2.25
	Ours w/ caption	3.00	5.00	1.67	1.67	2.83

C.4 Interpretation

The per-scenario table shows that improvements are not uniform across physical phenomena. Both of our complete configurations outperform both Wan baselines in aggregate, but the optional caption does not provide a monotonic gain. It raises CD only slightly and lowers the other three aggregate criteria, with five per-scenario improvements and eight declines relative to the no-caption setting.

D Authoring Ablations and Process Evidence

This appendix expands the two scene-authoring ablations summarized in Table 2. They measure different stages from the final-video comparison in Appendix C: acceptance records diagnose whether scene construction reaches a reviewer-approved rollout, while preview ratings assess the physical behavior visible in the simulation render. A complete preview can therefore receive a quality score even when its run does not reach acceptance within the allocated budget.

D.1 Structured Scene Representation

We compare direct Genesis code generation with our structured DSL on all 13 benchmark scenarios. Both arms use the same model, task description, execution budget, online execution and visual feedback, and acceptance criterion. Direct generation reaches acceptance in 7/13 scenarios, compared with 13/13 for the structured workflow. Independently, we score the last complete preview from each arm. The structured workflow increases the mean four-criterion score from 3.54 to 4.19, improving nine scenarios, tying two, and declining on two.

Table 7 separates these two outcomes. In particular, a score beside a budget-exhausted direct run describes its last complete preview; it does not reclassify the run as accepted.

Figure 5 illustrates one large positive difference. Both conditions retain execution and visual feedback. In the scored final direct-code preview, the initially recognizable colored pillows collapse into thin, sharply folded surfaces with severe volume loss. The structured workflow retains identifiable cushion bodies through compression, tumbling, and settling, consistent with preview-quality scores of 2.25 and 4.50, respectively.

D.2 Iterative Review and Refinement

Loop off stops after the initial rollout; loop on uses reviewer feedback until the final retained accepted rollout. Comparisons follow the same refinement lineage. Table 8 reports the recorded first-acceptance rounds used by the cumulative curve.

The initial endpoint reaches acceptance in 1/13 scenarios and produces a complete preview in 10/13. The final endpoint reaches acceptance and produces a complete preview in all 13. Three initial executions have no video, and the sand-rabbit initial video uses a below-table camera: the tabletop occludes the rabbit's collapse and the resulting sand pile, leaving insufficient visible evidence to

Table 7: **Structured-representation ablation by scenario.** Status is accepted within budget; scores evaluate the last complete preview.

Scenario	Direct status	Direct	DSL	Δ
Dominoes	Accepted	4.75	4.75	0.00
Two-ball collision	Accepted	4.50	4.50	0.00
Bowling	Accepted	4.75	3.50	-1.25
Newton's cradle	Budget exhausted	3.25	3.75	$+0.50$
Hanging lanterns	Budget exhausted	2.75	5.00	$+2.25$
Basketball bank shot	Accepted	4.00	4.50	$+0.50$
Pillows	Accepted	2.25	4.50	$+2.25$
Sand rabbit	Budget exhausted	3.75	4.75	$+1.00$
Champagne tower	Budget exhausted	2.75	4.00	$+1.25$
Water wheel	Budget exhausted	3.00	4.25	$+1.25$
Ball and mattress	Budget exhausted	2.75	3.25	$+0.50$
Vase fracture	Accepted	4.50	4.00	-0.50
Ball impacts glass tower	Accepted	3.00	3.75	$+0.75$
Mean	7/13 accepted	3.54	4.19	$+0.65$

score any of the four physical criteria. These four cases remain failures or unavailable observations in the success analysis and are not assigned quality score zero. Across the nine rateable endpoint pairs, quality increases from 3.50 to 4.25 ($+0.75$), with six improvements, two ties, and one decline.

Table 8: **Loop ablation by scenario.** Initial and final columns are simulation-preview means. A dash denotes no complete initial video. "Occluded view" means that the video is complete, but the below-table camera hides the sand-rabbit collapse and resulting sand pile, so the four physical criteria cannot be scored.

Scenario	Initial	Final	First accept round
Dominoes	3.50	4.75	2
Two-ball collision	2.00	4.50	4
Bowling	3.75	3.50	5
Newton's cradle	2.75	3.75	5
Hanging lanterns	5.00	5.00	2
Basketball bank shot	4.00	4.50	6
Pillows	–	4.50	9
Sand rabbit	Occluded view	4.75	2
Champagne tower	3.00	4.00	7
Water wheel	3.50	4.25	3
Ball and mattress	–	3.25	5
Vase fracture	4.00	4.00	1
Ball impacts glass tower	–	3.75	5
Matched mean ($n = 9$)	3.50	4.25	–

D.3 Basketball Refinement Trace

The basketball example illustrates how rollout evidence changes editable scene variables. Figure 6 shows the initial and accepted endpoints, and Table 9 summarizes selected intermediate diagnoses. In round 2 the ball leaves the valid simulation domain, so the rollout is rejected and refinement continues; this is an execution outcome, not a missing record.

Together, the structured-representation and loop ablations isolate complementary roles: the DSL improves the reliability and quality of scene specification, while iterative review repairs failures that remain after initial construction.

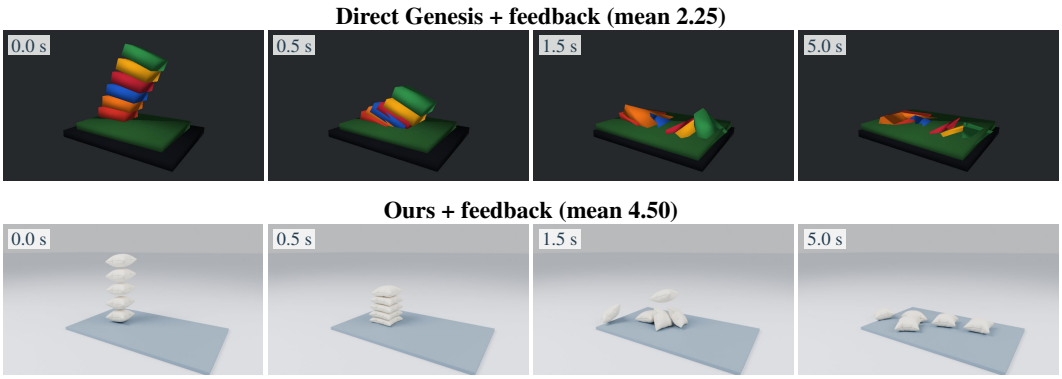


Figure 5: **Scene representation comparison.** Direct-code pillows (top) and the structured workflow (bottom), both with feedback. The bottom row uses a presentation rendering; the scores of 2.25 and 4.50 are measured on the original evaluation previews. Labels indicate playback time. Direct-code pillows lose volume and fold into thin surfaces, while the structured workflow retains cushion geometry.

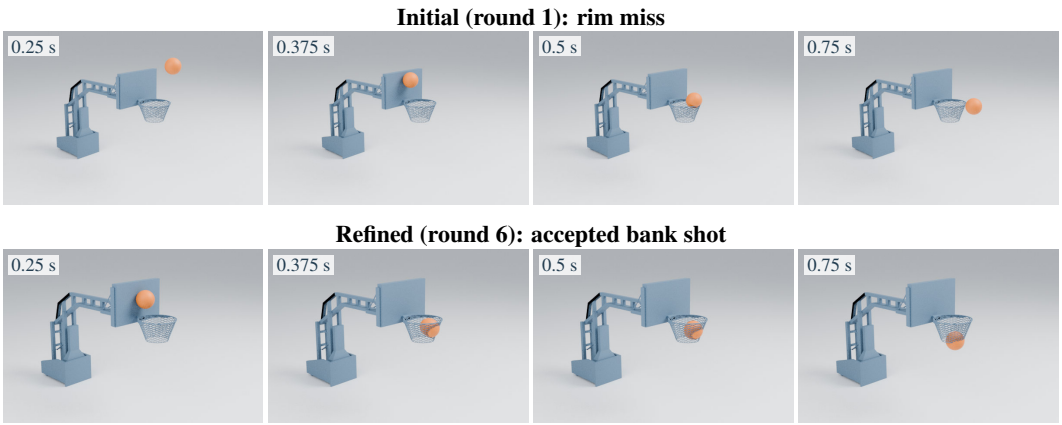


Figure 6: **Basketball refinement.** The initial (round 1) and accepted (round 6) rollouts preserve their recorded motion and per-round cameras; intermediate rounds are omitted.

Table 9: **Basketball refinement trace.** Selected milestones show the observed outcomes and corresponding review actions.

Round	Observed outcome	Review action
1	Ball contacts the backboard but misses the rim.	Request a launch correction.
2	Ball leaves the valid simulation domain during the rollout.	Reject the rollout and continue refinement.
3	The corrected trajectory remains too shallow.	Request a steeper downward launch.
4	Event readability remains inadequate.	Revise the launch and camera.
5	Ball reaches the support region without a clear accepted passage.	Revise hoop asset/orientation and launch state.
6	Backboard-to-rim passage and aftermath are visible.	Accept the rollout.

E Complete Visual Results

Each scenario shows Ours w/ caption, Ours w/o caption, Wan 2.2 I2V, and Wan 2.2 T2V. Scores average three GPT-5.6-Sol evaluations; Mean equally weights OP, CD, SG, and PR (1–5). Frames at 0, 0.5, 2, and 4 seconds come from the exact scored videos; evaluation uses the entire sequence. Selected observations are English translations from the first evaluation, not a consensus of the three calls.

E.1 Sand rabbit

Method	OP	CD	SG	PR	Mean
Ours w/ caption	5.00	5.00	5.00	3.33	4.58
Ours w/o caption	5.00	5.00	4.67	3.00	4.42
Wan 2.2 I2V	4.67	3.00	3.33	2.00	3.25
Wan 2.2 T2V	5.00	3.00	4.33	2.00	3.58

Ours w/ caption Mean 4.58



Ours w/o caption Mean 4.42



Wan 2.2 I2V Mean 3.25



Wan 2.2 T2V Mean 3.58



Selected evaluator observations. *Wan 2.2 T2V:* In frames 27–37 (1.69–2.31 s), the rabbit first crouches like a living animal, then melts like a continuous soft body, lacking local crumbling, granular sliding, and natural piling. *Ours w/ caption:* The stable loose pile with scattered boundary grains is plausible, but the main body melts almost as a whole in frames 1–4; the granular cascade is slightly too fast and smooth.

E.2 DOMINOES

Method	OP	CD	SG	PR	Mean
Ours w/ caption	4.00	4.67	4.67	3.33	4.17
Ours w/o caption	5.00	5.00	5.00	4.00	4.75
Wan 2.2 I2V	5.00	3.67	3.67	2.67	3.75
Wan 2.2 T2V	4.00	5.00	4.67	4.00	4.42

Ours w/ caption Mean 4.17



Ours w/o caption Mean 4.75



Wan 2.2 I2V Mean 3.75



Wan 2.2 T2V Mean 4.42



E.3 TWO-BALL COLLISION

Method	OP	CD	SG	PR	Mean
Ours w/ caption	5.00	5.00	4.00	3.67	4.42
Ours w/o caption	4.00	5.00	3.67	3.67	4.08
Wan 2.2 I2V	5.00	1.67	5.00	1.33	3.25
Wan 2.2 T2V	2.67	4.00	3.00	2.00	2.92

Ours w/ caption Mean 4.42



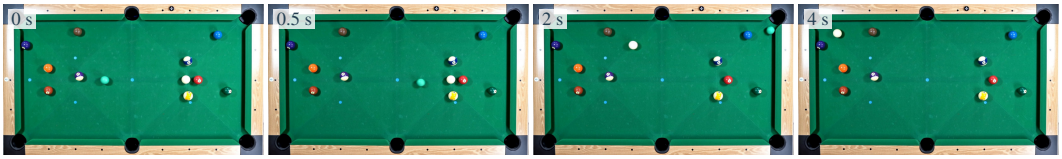
Ours w/o caption Mean 4.08



Wan 2.2 I2V Mean 3.25



Wan 2.2 T2V Mean 2.92



Selected evaluator observations. *Ours w/o caption:* Both white balls and the table remain intact in frames 0–80 without additions or losses, but an unrequested blue dot persists near the left ball's surface. *Ours w/ caption:* Two distinct balls and the intact table persist in frames 0–80, without disappearance, duplication, merging, or obvious deformation.

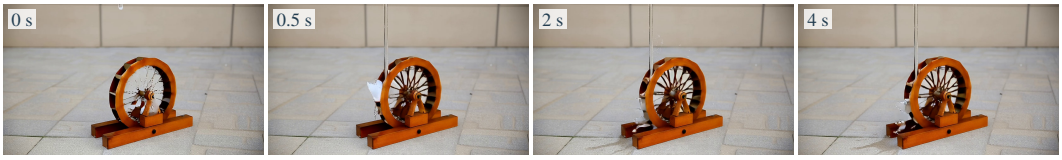
E.4 WATER WHEEL

Method	OP	CD	SG	PR	Mean
Ours w/ caption	3.33	4.00	4.33	2.67	3.58
Ours w/o caption	3.33	2.67	4.00	2.33	3.08
Wan 2.2 I2V	3.00	2.33	4.33	2.00	2.92
Wan 2.2 T2V	5.00	2.00	5.00	2.00	3.50

Ours w/ caption Mean 3.58



Ours w/o caption Mean 3.08



Wan 2.2 I2V Mean 2.92



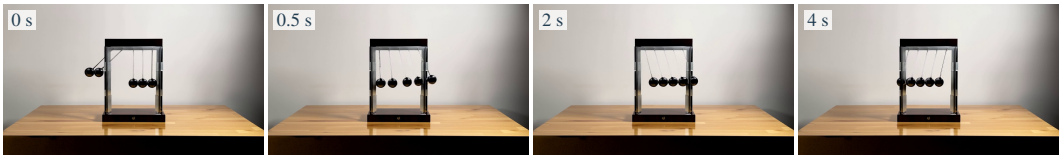
Wan 2.2 T2V Mean 3.50



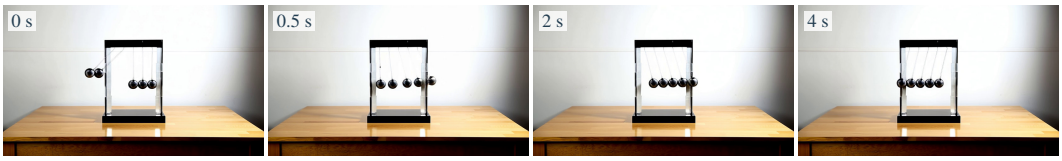
E.5 NEWTON'S CRADLE (TWO BALLS)

Method	OP	CD	SG	PR	Mean
Ours w/ caption	5.00	3.00	4.67	2.33	3.75
Ours w/o caption	5.00	3.00	5.00	2.33	3.83
Wan 2.2 I2V	3.00	4.00	3.67	3.00	3.42
Wan 2.2 T2V	4.00	2.67	4.00	2.00	3.17

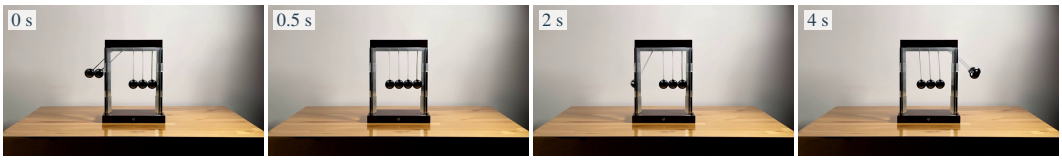
Ours w/ caption Mean 3.75



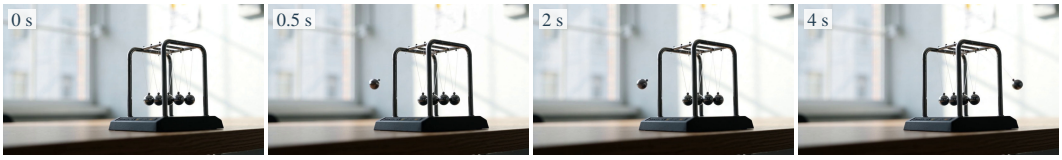
Ours w/o caption Mean 3.83



Wan 2.2 I2V Mean 3.42



Wan 2.2 T2V Mean 3.17

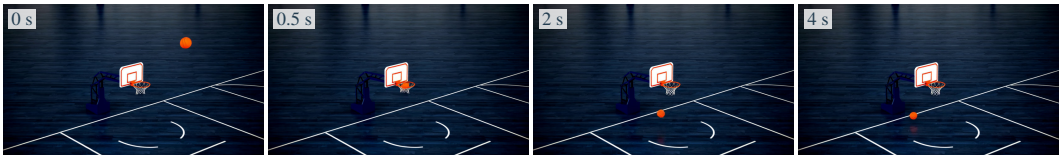


Selected evaluator observations. *Ours w/ caption:* Swinging and collisions are broadly continuous and damped, but at frames 10, 22, and later cycles several or all balls fan out together, departing from plausible Newton's-cradle momentum transfer.

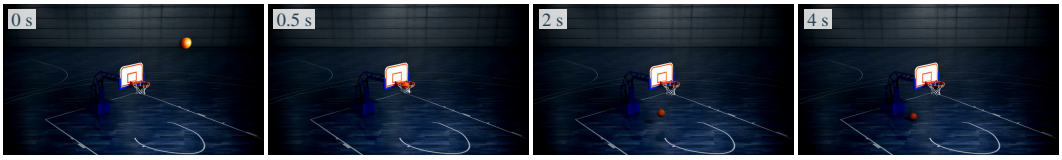
E.6 BASKETBALL BANK SHOT

Method	OP	CD	SG	PR	Mean
Ours w/ caption	4.67	4.67	4.33	3.33	4.25
Ours w/o caption	4.67	5.00	4.33	3.67	4.42
Wan 2.2 I2V	4.00	3.67	3.33	3.00	3.50
Wan 2.2 T2V	4.00	2.67	2.33	2.00	2.75

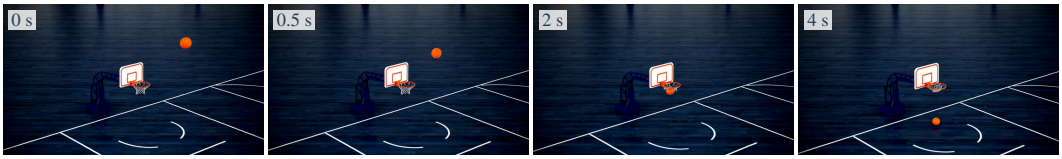
Ours w/ caption Mean 4.25



Ours w/o caption Mean 4.42



Wan 2.2 I2V Mean 3.50



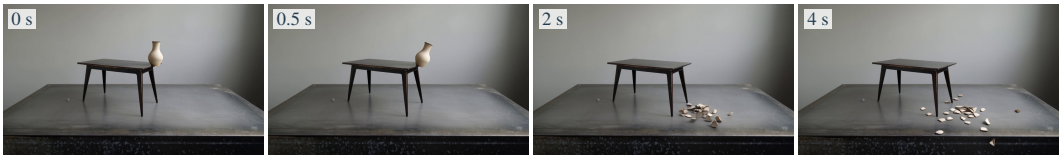
Wan 2.2 T2V Mean 2.75



E.7 VASE FRACTURE

Method	OP	CD	SG	PR	Mean
Ours w/ caption	3.33	3.33	3.67	2.00	3.08
Ours w/o caption	4.00	4.00	3.67	2.33	3.50
Wan 2.2 I2V	3.33	2.33	1.00	1.00	1.92
Wan 2.2 T2V	4.00	2.33	1.00	1.00	2.08

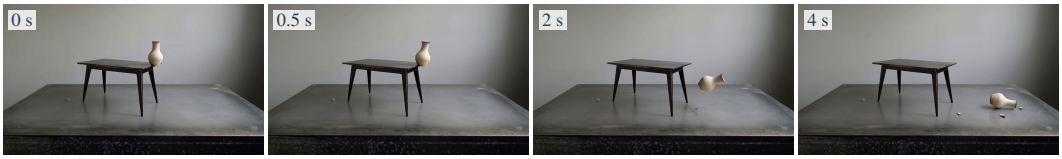
Ours w/ caption Mean 3.08



Ours w/o caption Mean 3.50



Wan 2.2 I2V Mean 1.92



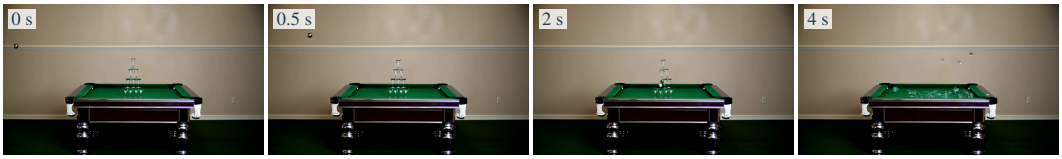
Wan 2.2 T2V Mean 2.08



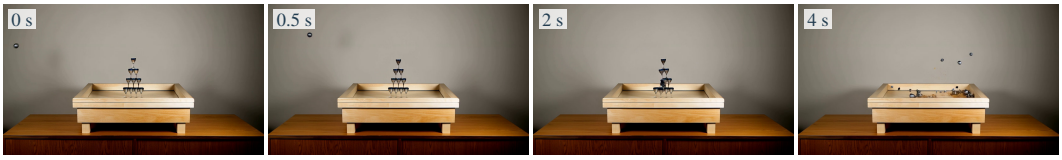
E.8 BALL IMPACTS GLASS TOWER

Method	OP	CD	SG	PR	Mean
Ours w/ caption	3.00	5.00	1.67	1.67	2.83
Ours w/o caption	2.00	4.33	1.33	1.33	2.25
Wan 2.2 I2V	3.33	4.00	2.00	2.00	2.83
Wan 2.2 T2V	3.67	2.00	2.33	1.67	2.42

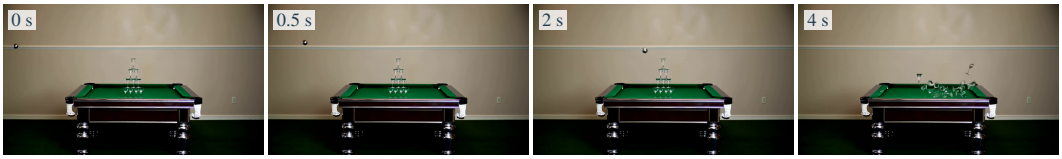
Ours w/ caption Mean 2.83



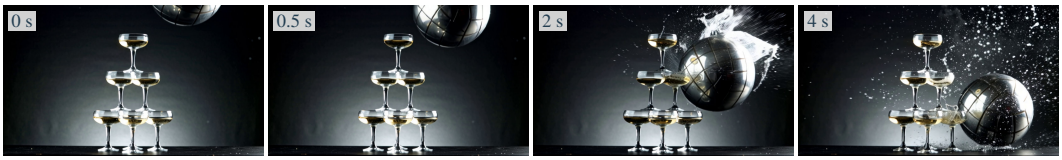
Ours w/o caption Mean 2.25



Wan 2.2 I2V Mean 2.83



Wan 2.2 T2V Mean 2.42



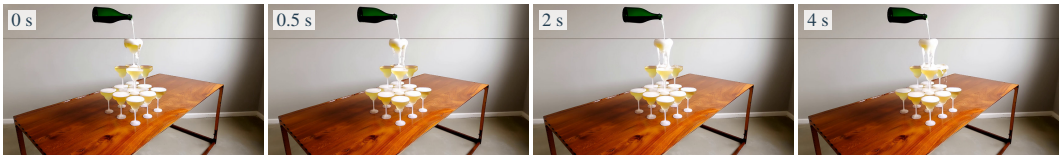
E.9 CHAMPAGNE TOWER

Method	OP	CD	SG	PR	Mean
Ours w/ caption	5.00	3.00	2.67	2.00	3.17
Ours w/o caption	4.33	2.33	3.00	2.00	2.92
Wan 2.2 I2V	5.00	2.33	3.00	2.00	3.08
Wan 2.2 T2V	4.00	2.67	4.00	3.00	3.42

Ours w/ caption Mean 3.17



Ours w/o caption Mean 2.92



Wan 2.2 I2V Mean 3.08



Wan 2.2 T2V Mean 3.42



Selected evaluator observations. *Wan 2.2 T2V:* Continuous pouring, foaming, and downward overflow are broadly understandable, but several streams remain nearly fixed thin lines bypassing glass interiors, with weak changes in liquid levels and receiving responses. *Ours w/ caption:* Most liquid levels barely change during pouring; thick, nearly static white columns connect the top and middle tiers, while most bottom glasses lack corresponding inflow or overflow, making fluid response implausible.

E.10 PILLOWS

Method	OP	CD	SG	PR	Mean
Ours w/ caption	4.67	3.00	2.00	2.00	2.92
Ours w/o caption	4.67	3.00	2.00	3.00	3.17
Wan 2.2 I2V	2.33	4.00	2.67	2.00	2.75
Wan 2.2 T2V	4.33	1.00	2.33	2.00	2.42

Ours w/ caption Mean 2.92



Ours w/o caption Mean 3.17



Wan 2.2 I2V Mean 2.75



Wan 2.2 T2V Mean 2.42

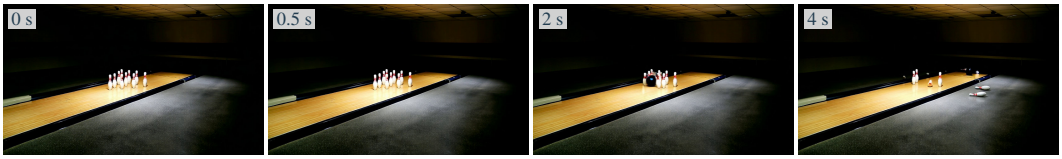


Selected evaluator observations. *Ours w/ caption:* In frames 0–24, several separated, unsupported pillows remain suspended; upper pillows stay airborne even after lower ones have collapsed.

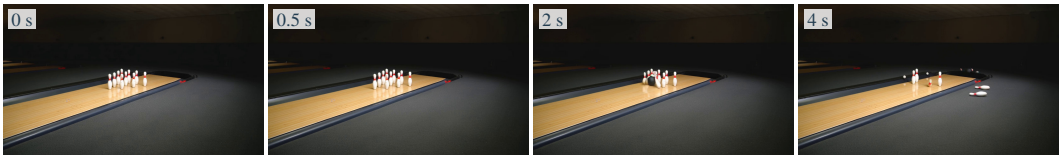
E.11 BOWLING

Method	OP	CD	SG	PR	Mean
Ours w/ caption	2.00	3.00	2.00	2.00	2.25
Ours w/o caption	2.00	3.00	2.33	2.00	2.33
Wan 2.2 I2V	1.00	2.00	1.33	1.00	1.33
Wan 2.2 T2V	3.67	2.33	2.00	2.00	2.50

Ours w/ caption Mean 2.25



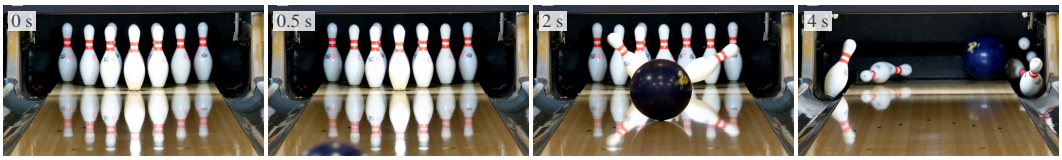
Ours w/o caption Mean 2.33



Wan 2.2 I2V Mean 1.33



Wan 2.2 T2V Mean 2.50



E.12 BALL AND MATTRESS

Method	OP	CD	SG	PR	Mean
Ours w/ caption	5.00	5.00	3.00	2.33	3.83
Ours w/o caption	5.00	5.00	4.33	4.00	4.58
Wan 2.2 I2V	4.67	4.67	4.67	4.00	4.50
Wan 2.2 T2V	4.33	2.33	4.67	2.67	3.50

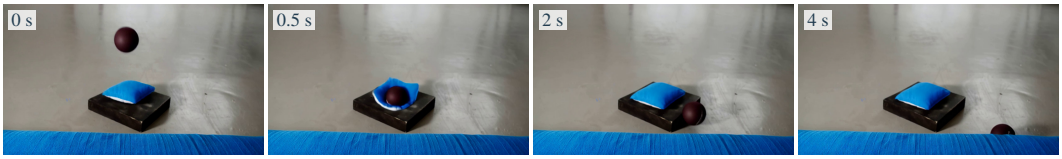
Ours w/ caption Mean 3.83



Ours w/o caption Mean 4.58



Wan 2.2 I2V Mean 4.50



Wan 2.2 T2V Mean 3.50



E.13 HANGING LANTERNS

Method	OP	CD	SG	PR	Mean
Ours w/ caption	3.00	3.67	4.00	3.00	3.42
Ours w/o caption	5.00	4.33	5.00	4.00	4.58
Wan 2.2 I2V	3.33	4.00	4.00	3.33	3.67
Wan 2.2 T2V	5.00	3.00	5.00	3.67	4.17

Ours w/ caption Mean 3.42



Ours w/o caption Mean 4.58



Wan 2.2 I2V Mean 3.67



Wan 2.2 T2V Mean 4.17

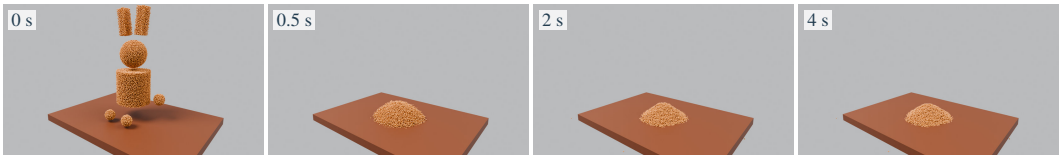


E.14 AUTHORING ABLATIONS: SAND RABBIT

These are simulation previews, evaluated once per unique video by GPT-5.6-Sol with high reasoning effort. Direct Genesis retains feedback and uses its last complete preview; Structured DSL uses the structured workflow. Loop off and Loop on denote the original first attempt and the final preview, respectively. A shared final video is labelled Structured DSL / Loop on. Where cached geometry is available, keyframes are re-rendered with flat object colors and studio lighting, preserving geometry, cameras, and frame times; other keyframes remain original. Scores refer to the original videos. Quality averages OP, CD, SG, and PR; execution failures and unobservable evidence receive no quality score. Selected observations condense the recorded evaluator rationales.

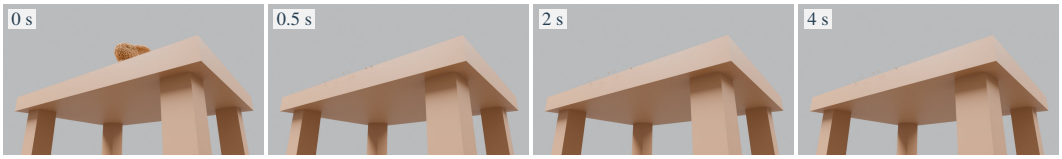
Condition	OP	CD	SG	PR	Mean
Direct Genesis	4.00	5.00	3.00	3.00	3.75
Loop off (occluded)	–	–	–	–	–
Structured DSL / Loop on	5.00	5.00	5.00	4.00	4.75

Direct Genesis Mean 3.75



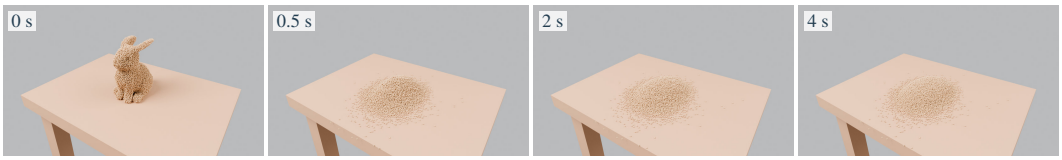
Selected observation: Collapse, outward particle flow, and settling are present, but disappearing clusters and cohesive head and ear blocks make the response overly segmented.

Loop off (occluded) Not rateable



Selected observation: Disintegration, granular flow, piling, and stopping are occluded, leaving insufficient evidence to assess material response.

Structured DSL / Loop on Mean 4.75

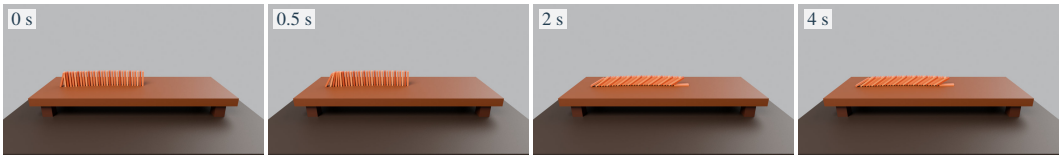


Selected observation: Granular collapse and stable piling are broadly plausible, but early whole-body flattening and briefly cohesive ears provide limited progressive fine-grain disintegration.

E.15 AUTHORING ABLATIONS: DOMINOES

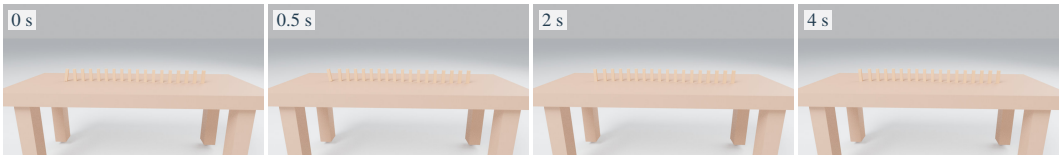
Condition	OP	CD	SG	PR	Mean
Direct Genesis	4.00	5.00	5.00	5.00	4.75
Loop off	5.00	1.00	5.00	3.00	3.50
Structured DSL / Loop on	5.00	5.00	5.00	4.00	4.75

Direct Genesis Mean 4.75



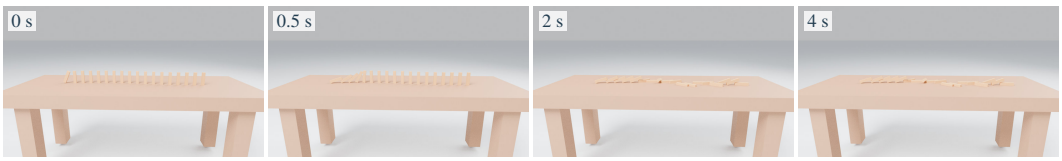
Selected observation: The pieces remain approximately rigid and transmit toppling through successive collisions, with a clear propagation front and stable final settling.

Loop off Mean 3.50



Selected observation: The first domino only rocks slightly and does not topple its neighbor; the remaining row stays upright throughout.

Structured DSL / Loop on Mean 4.75

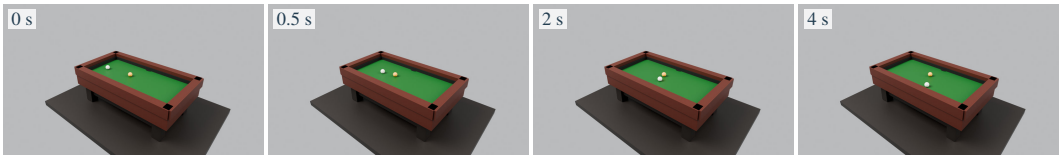


Selected observation: The chain reaction and settling are broadly plausible, but some fallen pieces scatter sideways and reorient more than a typical closely spaced domino chain.

E.16 AUTHORING ABLATIONS: TWO-BALL COLLISION

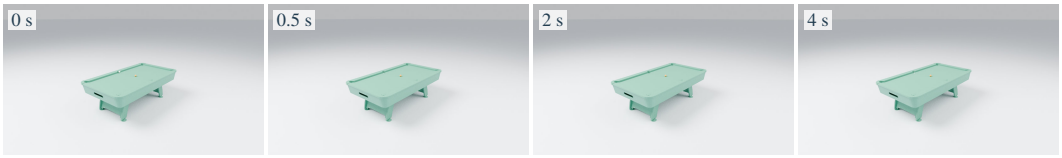
Condition	OP	CD	SG	PR	Mean
Direct Genesis	3.00	5.00	5.00	5.00	4.50
Loop off	2.00	1.00	4.00	1.00	2.00
Structured DSL / Loop on	5.00	5.00	4.00	4.00	4.50

Direct Genesis Mean 4.50



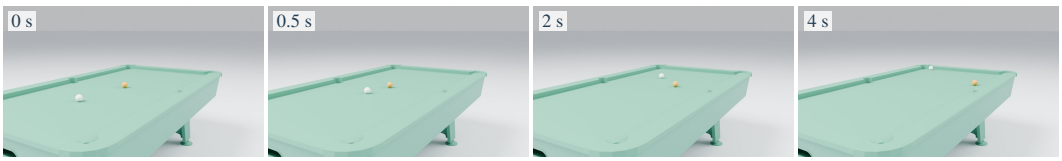
Selected observation: The yellow ball gains motion while the white ball deflects and retains velocity, showing plausible qualitative momentum transfer and subsequent slowing.

Loop off Mean 2.00



Selected observation: The sequence is nearly static, with no visible approach, off-center collision, momentum transfer, or two-ball separation.

Structured DSL / Loop on Mean 4.50



Selected observation: Post-impact transfer and diverging trajectories are plausible, but the late response near the corner pocket lacks a convincing drop, rebound, or stop.

E.17 Authoring ablations: Water wheel

Loop off uses the first preview of the inherited parent run. Loop on uses the final rerun, which retains the accepted scene parameters and adds an outflow boundary.

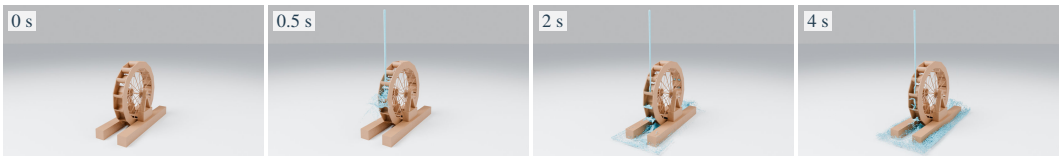
Condition	OP	CD	SG	PR	Mean
Direct Genesis	5.00	1.00	4.00	2.00	3.00
Loop off	5.00	2.00	5.00	2.00	3.50
Structured DSL / Loop on	5.00	5.00	4.00	3.00	4.25

Direct Genesis Mean 3.00



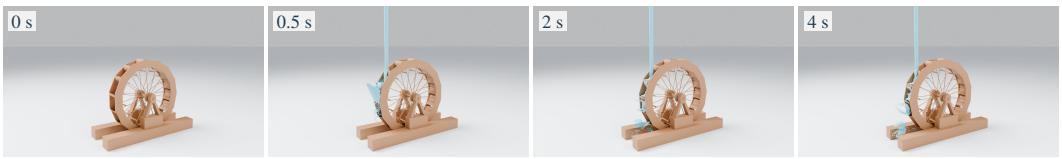
Selected observation: Water strikes the upper support beam instead of the wheel paddles, and the wheel remains nearly stationary, missing the requested sustained rotation.

Loop off Mean 3.50



Selected observation: Water reaches the wheel, but continuous rotation is not visually evident; sparse particle-like splashes weaken the fluid response.

Structured DSL / Loop on Mean 4.25

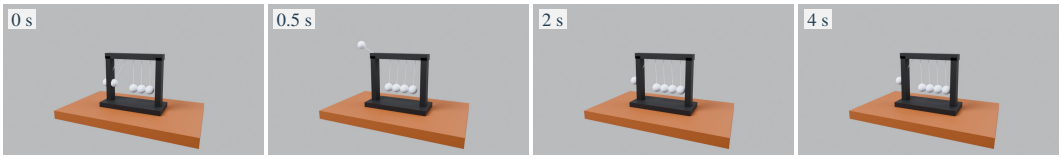


Selected observation: Water drives wheel rotation, but early sheet-like splashes and later sparse particles weaken fluid continuity, merging, and natural ground dissipation.

E.18 AUTHORING ABLATIONS: NEWTON'S CRADLE (TWO BALLS)

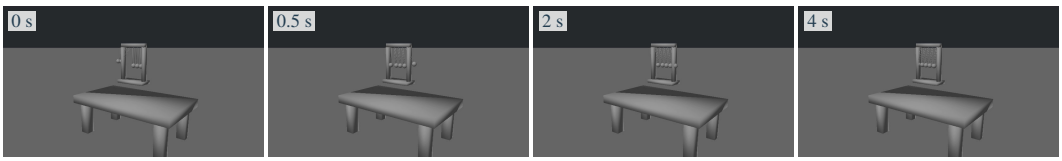
Condition	OP	CD	SG	PR	Mean
Direct Genesis	5.00	1.00	5.00	2.00	3.25
Loop off	5.00	2.00	2.00	2.00	2.75
Structured DSL / Loop on	5.00	3.00	5.00	2.00	3.75

Direct Genesis Mean 3.25



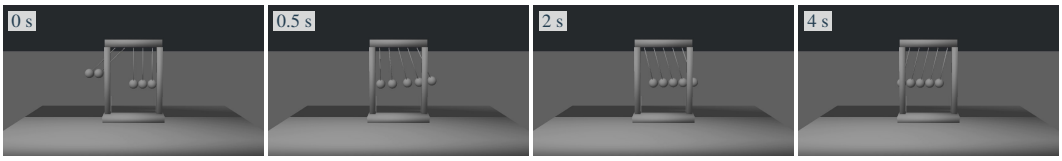
Selected observation: Although two left balls start raised, only one subsequently swings clearly; the two rightmost balls never swing out together or alternate with the left pair.

Loop off Mean 2.75



Selected observation: Rigid pendular motion is present, but a two-ball input produces mainly one-ball output, violating the intended qualitative collision response.

Structured DSL / Loop on Mean 3.75

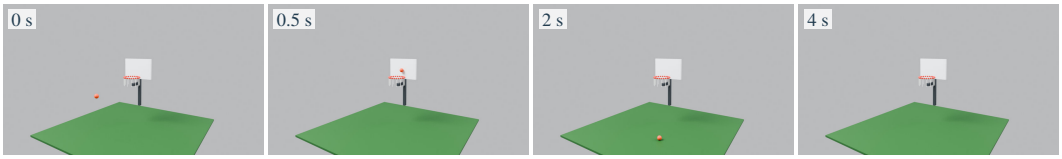


Selected observation: Initial transfer occurs, but later four or five balls swing together instead of maintaining the expected two-in, two-out collision response.

E.19 AUTHORING ABLATIONS: BASKETBALL BANK SHOT

Condition	OP	CD	SG	PR	Mean
Direct Genesis	5.00	4.00	4.00	3.00	4.00
Loop off	5.00	2.00	5.00	4.00	4.00
Structured DSL / Loop on	5.00	5.00	4.00	4.00	4.50

Direct Genesis Mean 4.00



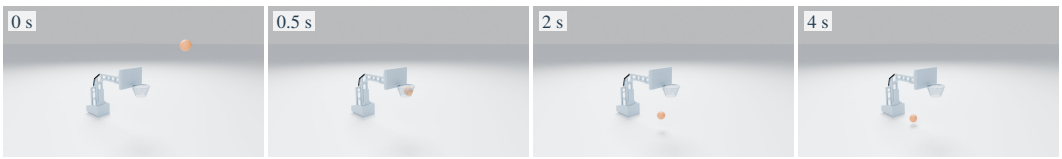
Selected observation: The trajectory remains nearly parabolic near the board rather than showing a clear collision deflection, and the net barely responds to passage.

Loop off Mean 4.00



Selected observation: The change in horizontal direction near the board is consistent with rebound and gravity, although the exact contact is visually ambiguous.

Structured DSL / Loop on Mean 4.50



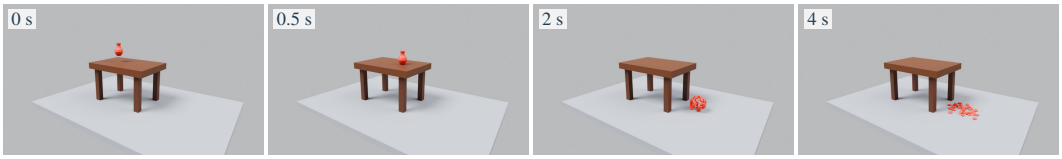
Selected observation: Board deflection, hoop passage, and damped bouncing are broadly reasonable, but the net deforms and follows the ball only weakly.

E.20 AUTHORING ABLATIONS: VASE FRACTURE

The initial and final endpoints use the same recorded preview. It is shown only under Structured DSL / Loop on; the identical Loop off score is retained in the table.

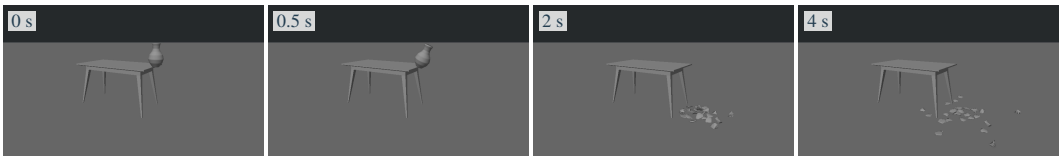
Condition	OP	CD	SG	PR	Mean
Direct Genesis	5.00	5.00	5.00	3.00	4.50
Loop off	5.00	4.00	3.00	4.00	4.00
Structured DSL / Loop on	5.00	4.00	3.00	4.00	4.00

Direct Genesis Mean 4.50



Selected observation: Tipping, falling, and later settling are broadly plausible, but the immediate debris resembles a dense cluster of thick blocks rather than thin ceramic shards.

Structured DSL / Loop on Mean 4.00



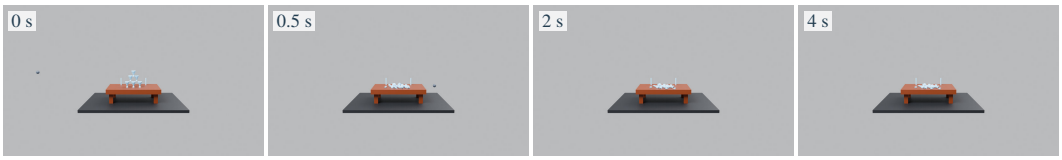
Selected observation: Cracking, rigid fragment separation, and scattering resemble ceramic fracture, although some fragments keep making small late jumps and drifts.

E.21 Authoring ablations: Ball impacts glass tower

Loop off: no video (execution failure). The first three attempts fail before simulation according to the supplied records; no original error log is available. The first complete preview is round four. This later preview is not substituted for the Loop off condition.

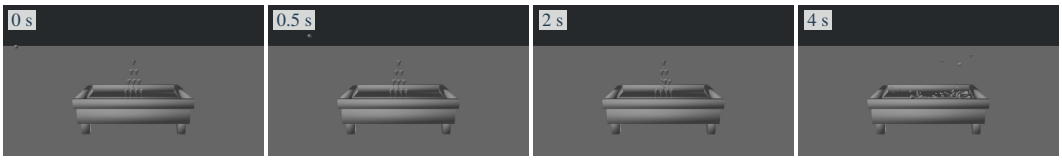
Condition	OP	CD	SG	PR	Mean
Direct Genesis	5.00	1.00	5.00	1.00	3.00
Loop off (no video)	–	–	–	–	–
Structured DSL / Loop on	5.00	5.00	2.00	3.00	3.75
First available after failure	3.00	4.00	4.00	3.00	3.50

Direct Genesis Mean 3.00



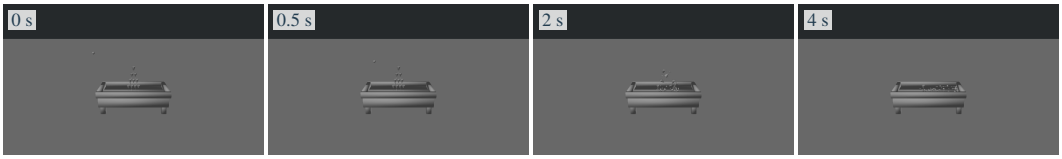
Selected observation: The tower collapses while the ball is still distant and becomes rubble before the ball arrives, reversing the required causal sequence.

Structured DSL / Loop on Mean 3.75



Selected observation: Initial local fracture and cascading collapse are plausible, but later fragments repeatedly accelerate and reverse vertically in midair, weakening the settling response.

First available after failure Mean 3.50



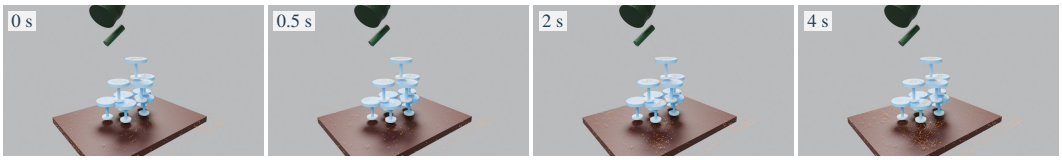
Selected observation: Fracture, scattering, and settling are broadly plausible, but the ball's post-impact response and clear momentum transfer are missing.

E.22 AUTHORING ABLATIONS: CHAMPAGNE TOWER

The DSL comparison uses the bottle-pouring task; the loop comparison uses the separate seven-round glass-pouring run. Direct code is scored on round six, its last complete preview; round seven timed out and round eight was cancelled.

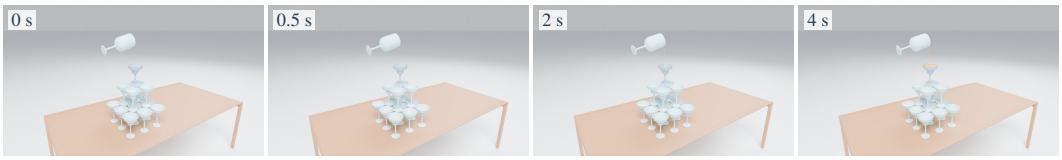
Condition	OP	CD	SG	PR	Mean
Direct Genesis	3.00	3.00	3.00	2.00	2.75
Loop off	5.00	2.00	3.00	2.00	3.00
Structured DSL	5.00	4.00	3.00	4.00	4.00
Loop on (glass-pouring run)	5.00	5.00	4.00	2.00	4.00

Direct Genesis Mean 2.75



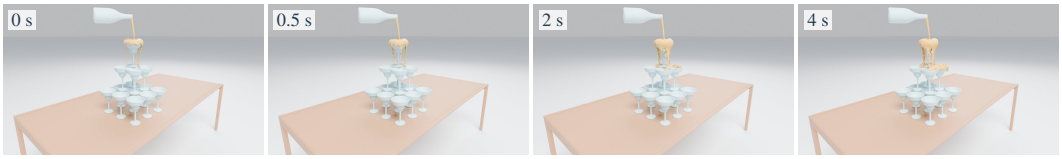
Selected observation: Sparse particles contact glasses and the table but lack continuous flow, accumulation, overflow sheets, and merging expected of liquid.

Loop off Mean 3.00



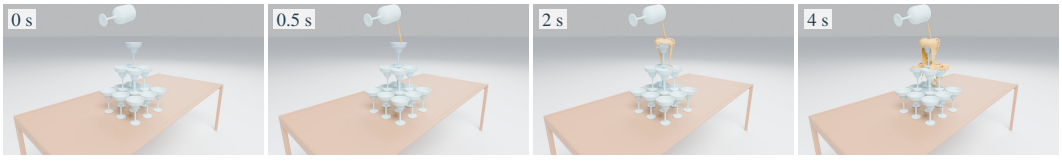
Selected observation: The liquid accumulates as a growing rigid spike or granular pile rather than spreading continuously over the rim into lower glasses.

Structured DSL Mean 4.00



Selected observation: The stream divides into sheets and droplets at glass rims with broadly plausible flow, although some motion is sparse, granular, or escapes between glasses.

Loop on (glass-pouring run) Mean 4.00



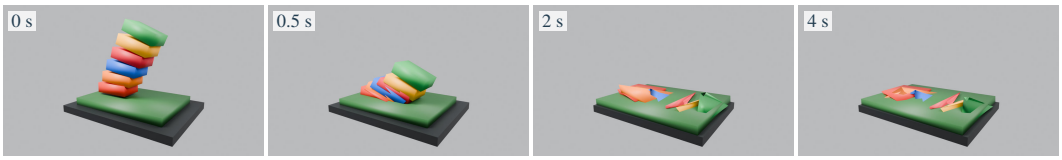
Selected observation: Inflow and cascading occur, but a large dome above the top rim and thick rope-like overflow resemble viscous clumps rather than low-viscosity wine.

E.23 AUTHORING ABLATIONS: PILLOWS

Loop off: no video (execution failure). The first attempt fails a contact-memory budget check before producing a complete preview. The first available preview is round three. This later preview is not substituted for the Loop off condition.

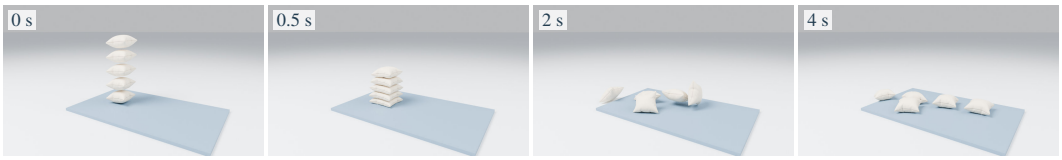
Condition	OP	CD	SG	PR	Mean
Direct Genesis	2.00	3.00	3.00	1.00	2.25
Loop off (no video)	–	–	–	–	–
Structured DSL / Loop on	5.00	4.00	5.00	4.00	4.50
First available after failure	4.00	3.00	4.00	3.00	3.50

Direct Genesis Mean 2.25



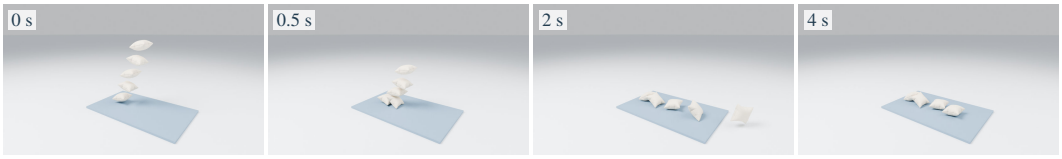
Selected observation: The cushions lose volume and collapse into sharp, paper-like polygons instead of showing thick soft-body compression, bending, and recovery.

Structured DSL / Loop on Mean 4.50



Selected observation: Compression, bending, tumbling, and dissipation are visible, but brief sharp-corner support and rebound appear slightly too stiff and elastic.

First available after failure Mean 3.50

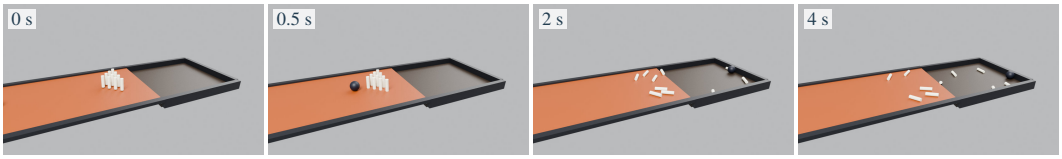


Selected observation: Compression and damping are visible, but one pillow rebounds far sideways with prolonged tumbling, and some sharp folding is inconsistent with soft dissipative cushions.

E.24 AUTHORING ABLATIONS: BOWLING

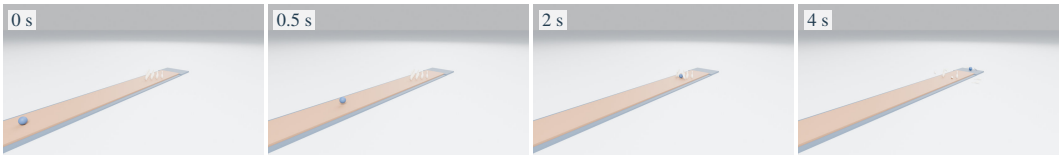
Condition	OP	CD	SG	PR	Mean
Direct Genesis	5.00	5.00	5.00	4.00	4.75
Loop off	4.00	3.00	4.00	4.00	3.75
Structured DSL / Loop on	4.00	2.00	4.00	4.00	3.50

Direct Genesis Mean 4.75



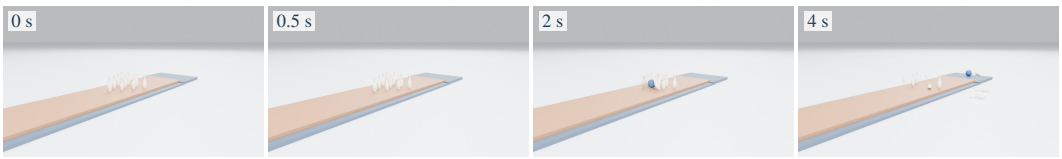
Selected observation: Chain collisions, rotation, and scattering show plausible momentum transfer, although some pins are thrown upward with slightly excessive elasticity.

Loop off Mean 3.75



Selected observation: Toppling, pin-to-pin collisions, and lateral scattering are continuous and broadly plausible; detailed late rolling and settling are difficult to discern.

Structured DSL / Loop on Mean 3.50



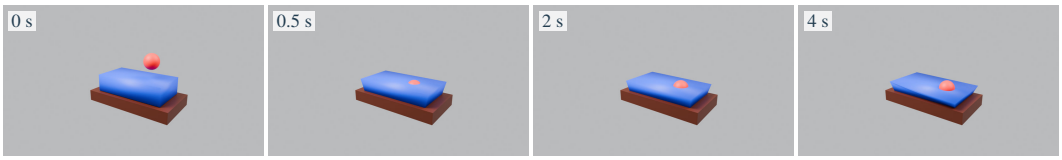
Selected observation: Collision-driven toppling and continued ball motion are reasonable, but some fallen pins slide for too long with weak rolling and damping.

E.25 Authoring ablations: Ball and mattress

Loop off: no video (execution failure). The first attempt fails a particle-count budget check before producing a complete preview. The first available preview is round two. This later preview is not substituted for the Loop off condition.

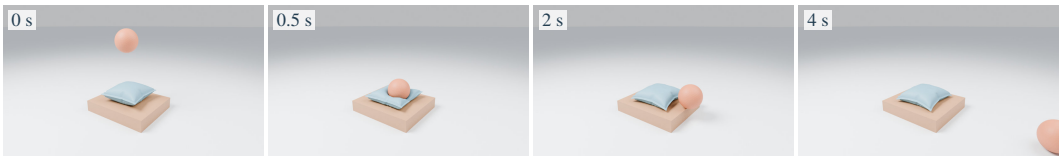
Condition	OP	CD	SG	PR	Mean
Direct Genesis	4.00	2.00	3.00	2.00	2.75
Loop off (no video)	–	–	–	–	–
Structured DSL / Loop on	5.00	2.00	3.00	3.00	3.25
First available after failure	5.00	3.00	2.00	2.00	3.00

Direct Genesis Mean 2.75



Selected observation: The mattress oscillates in overall thickness rather than forming a local indentation; repeated sinking and emergence show little damping and no plausible transition to rolling.

Structured DSL / Loop on Mean 3.25



Selected observation: The upper cushion indents and recovers, but the actual mattress barely deforms under the heavy ball, conflicting with the requested material response.

First available after failure Mean 3.00



Selected observation: Bounce heights diminish, but the soft mattress never compresses and the ball rebounds through a rigid plane, giving the wrong material response.

E.26 AUTHORING ABLATIONS: HANGING LANTERNS

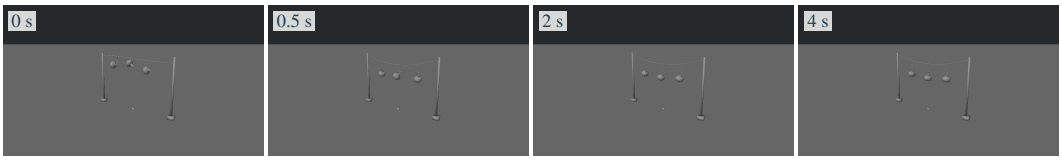
Condition	OP	CD	SG	PR	Mean
Direct Genesis	5.00	2.00	2.00	2.00	2.75
Loop off	5.00	5.00	5.00	5.00	5.00
Structured DSL / Loop on	5.00	5.00	5.00	5.00	5.00

Direct Genesis Mean 2.75



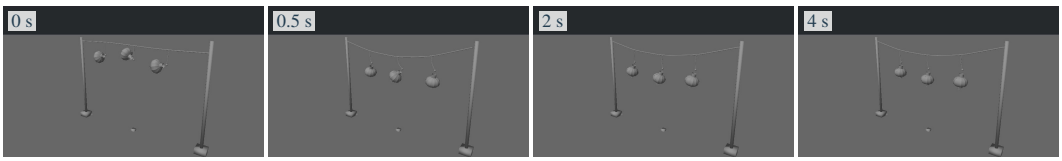
Selected observation: The rope makes an excessive whip-like motion and stops rapidly; lanterns largely follow it instead of showing clear, gradually damped independent swinging.

Loop off Mean 5.00



Selected observation: Rope and lantern motion is coupled and continuous, with alternating swings and diminishing amplitude that naturally approach rest.

Structured DSL / Loop on Mean 5.00



Selected observation: The lanterns move continuously like damped pendulums while the rope changes curvature, settling smoothly without teleportation or unexplained acceleration.