VIDPHYSORACLE: BENCHMARKING VIDEO WORLD MODELS AGAINST PHYSICALLY GROUNDED FUTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent video generation and world models can synthesize visually compelling futures, yet their ability to capture real-world physics remains unclear. Existing physics evaluations largely rely on hand-crafted rules or learned judges, requiring the evaluator to infer what should physically happen. We introduce VidPhysOracle, a benchmark and physically grounded video dataset that evaluates generated futures against oracle trajectories produced by physical simulation. Rather than asking whether a video merely looks physically plausible, VidPhysOracle measures how closely its dynamics match a known physical evolution. The dataset covers diverse rigid-body, deformable/soft-body, and fluid systems with controlled initial conditions and corresponding simulated futures. To reduce the simulation-to-real visual gap, we construct physics-preserving photorealistic videos through appearance transformation while retaining the underlying geometry and motion. We further introduce a reference-based unified physical consistency score to quantify dynamic agreement beyond pixel-level similarity, and benchmark seven representative open-source video and world models under identical initial conditions. Beyond evaluation, VidPhysOracle provides physically valid future trajectories for physics-aware fine-tuning and post-training, enabling existing video models to improve physical consistency. Overall, VidPhysOracle offers a unified framework for evaluating, diagnosing, and improving the physical capabilities of video world models.

Table 1: **Selected benchmark coverage.** ✓ denotes explicit coverage or use of the listed property; × denotes its absence; and / indicates coverage not specified under this material taxonomy.

Benchmark	Evaluation setup		Material coverage				Metric
	Physical consistency	Reference video	Rigid	Soft	Liquid	Gas	Rule-based metric
VBench (Huang et al., 2023)	×	×	/	/	/	/	×
VBench-2.0 (Zheng et al., 2025)	✓	×	/	/	/	/	×
VideoPhy (Bansal et al., 2024)	✓	×	✓	✓	✓	×	×
PhyGenBench (Meng et al., 2024)	✓	×	✓	✓	✓	×	×
Physion-Eval (Zhang et al., 2026)	✓	✓	✓	✓	✓	✓	×
Physics-IQ (Motamed et al., 2026)	✓	✓	✓	✓	✓	×	✓
RigidBench (Jain & Wu, 2026)	✓	✓	✓	×	×	×	✓
VidPhysOracle (ours)	✓	✓	✓	✓	✓	✓	✓

1 INTRODUCTION

Video generation is increasingly explored as a foundation for physical world modeling, including embodied learning, policy evaluation, and predictive simulation (Ali et al., 2025). Yet visual realism does not establish accurate physical prediction (Bansal et al., 2026). An object may bounce at the wrong time, or a convincing liquid stream may spread incorrectly. For a world model, the central question is therefore not whether a future looks plausible, but whether it follows the physical

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.

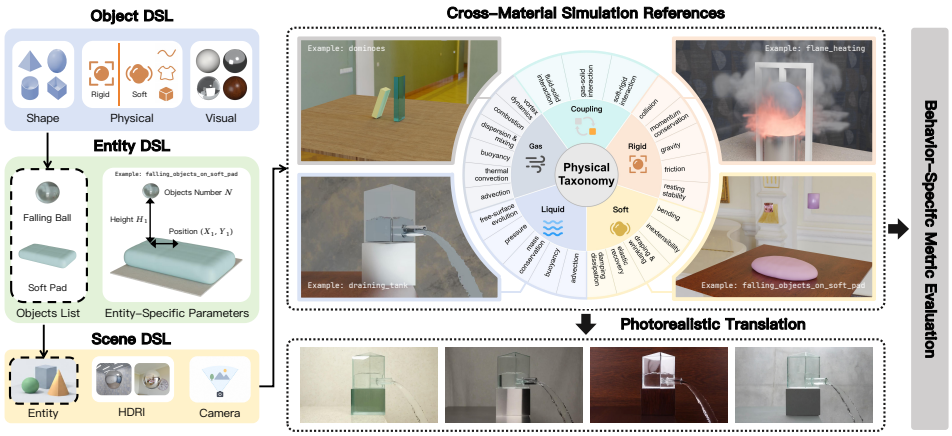


Figure 1: **Benchmark overview.** The benchmark comprises three stages. *Left:* a three-layer Domain-Specific Language (DSL) separates reusable object templates, scenario composition, and scene-level rendering conditions. *Middle:* a physical taxonomy organizes behaviours into four material categories and their cross-material coupling, with representative scenarios and photorealistic appearance translation. *Right:* the resulting reference data support behaviour-specific evaluation with physics-grounded metrics.

evolution implied by the underlying process. This distinction separates qualitative plausibility from fidelity to a specified physical process.

Existing physical evaluations mainly assess generated videos through explicit physical constraints or learned judgments. Physics-informed methods recover states or trajectories and test governing rules (Zhang et al., 2025), while learned evaluators judge whether generated behavior appears physically reasonable (Meng et al., 2024); however, both rely on the evaluator to infer the expected behavior and may miss or mislocalize physical errors (Zhang et al., 2026). This limitation has motivated reference-based evaluation: Physics-IQ compares predictions with real-world experimental observations (Motamed et al., 2026), and RigidBench provides simulated rigid-body rollouts (Jain & Wu, 2026). Yet existing reference-based benchmarks remain limited in physical coverage, and extending them beyond relatively structured rigid-body motion raises a more fundamental challenge: different materials expose physical errors through fundamentally different forms of motion, deformation, and flow.

Our key observation is that physical evaluation becomes much better defined once the expected future is made explicit. Rather than asking an evaluator to infer whether a generated motion is physically plausible, we can first construct a physically valid evolution under controlled initial conditions and then measure how the generated future deviates from it. This reframes physical evaluation from an open-ended judgment problem into a reference-grounded comparison problem. The remaining challenge is then not to invent a universal notion of similarity, but to identify the appropriate observable for each physical regime: object motion for rigid bodies, deformation for soft bodies, and evolving spatial support for fluids. This perspective suggests a unified principle for cross-material evaluation: *ground the target evolution with simulation, while adapting the measurement to the physics being observed.*

Following this principle, we introduce **VidPhysOracle**, a simulation-grounded benchmark that evaluates generated futures against oracle physical evolutions under controlled initial conditions, as illustrated in Figure 1. For each instance, a material-appropriate simulator produces the reference rollout, while the video model predicts its own future from the corresponding initial observation, enabling direct comparison against a known physical process. To move beyond a small set of handcrafted or rigid-body cases, we systematically construct 960 randomized scenes across twelve families spanning rigid bodies, soft bodies, liquids, gases, and cross-material interactions, with a hierarchical domain-specific language supporting controlled variation of physical configurations and scene appearance. We further employ a structure-conditioned video transfer model to reduce synthetic visual bias while preserving the geometry, motion, and underlying dynamics of the simulated reference. This construction provides diverse yet physically controlled futures for both reference-based evaluation and physics-aware model adaptation.

Our evaluation framework matches observables to material dynamics: object geometry and surface-point motion for rigid bodies, projected deformation response for soft bodies, and visible-support distribution and change for liquids and smoke. For continuous media, we extend multiscale occupancy comparison to temporal support changes, distinguishing spatial overlap from actual support evolution and penalizing frozen predictions when the reference changes. A fixed spatial mapping and frame-rate-based correspondence preserve position and timing errors without per-frame registration or temporal warping. These material-aware measurements provide an interpretable assessment of physical dynamics while retaining a unified reference-based evaluation protocol across heterogeneous systems.

Our contributions are threefold:

- **A cross-material physical benchmark.** We construct controlled simulation references across twelve scenario families spanning rigid-body, soft-body, liquid, gas, and cross-material dynamics, enabling systematic evaluation beyond predominantly rigid-body settings.
- **A dataset for evaluation and adaptation.** VidPhysOracle provides randomized physical scenes, simulation-derived annotations, and physics-preserving photorealistic variants, supporting both reference-based benchmarking and physics-aware model adaptation.
- **Material-aware reference-based evaluation.** We design geometry-, deformation-, and temporal-support-based measurements that capture characteristic physical errors across different material regimes under a unified oracle-grounded evaluation protocol.

2 VIDPHYSORACLE: BENCHMARK DESIGN AND CONSTRUCTION

VidPhysOracle evaluates physical prediction through controlled scenarios and simulation-derived reference rollouts. Its construction connects material-specific physical targets, systematic scene diversification, and evaluation-oriented annotations. These rollouts support both reference-based evaluation and physics-aware fine-tuning of video world models.

2.1 DESIGN PRINCIPLES AND PHYSICAL TAXONOMY

Our two-level taxonomy maps four material groups—rigid bodies, soft bodies, liquids, and gases—to their characteristic physical behaviours (Figure 1, middle). These targets guide scenario design and the corresponding behaviour-specific evaluation. We instantiate twelve scenario families, three per group: rigid-body motion and contact; deformation of ropes, cloth, and volumetric bodies; liquid transport and free-surface evolution; and gas transport, thermal effects, and vortex dynamics. A coupling branch captures liquid–solid, gas–solid, and soft–rigid interactions within these scenarios. The complete suite is given in Appendix C.1.

2.2 AGENTIC SCENARIO DIVERSIFICATION

An agent designs the scenario families around three requirements: controllable initial states, identifiable physical behaviours, and appearance variation independent of physical parameters. A three-layer *domain-specific language* (DSL) enables controlled and diverse scenario randomization (Figure 1, left). The *Object DSL* defines reusable object-level geometry and physical/visual materials; the *Entity DSL* specifies each scenario's object composition and initial layout while exposing entity-specific parameters for randomization; and the *Scene DSL* controls the environment and camera. For example, in the *falling objects on soft pad* scenario, the Entity DSL specifies the falling objects and soft pad as the basic scene composition, while their height and position are independently randomized across instances. The three layers are decoupled, allowing object-level properties, scenario evolution, and global visual appearance to be randomized independently. All scenarios start from rest and evolve naturally under gravity alone. within each family, solver settings remain fixed while designated parameters vary. This separation isolates physical variation from appearance and numerical configuration, enabling systematic scenario diversification while preserving the intended physical behaviour (Appendix C.2).

2.3 PHYSICS-GROUNDED ORACLE ROLLOUT GENERATION

Each scenario is simulated with a material-appropriate solver, configured to prioritize numerical stability and high-fidelity physical evolution over real-time performance. Rigid bodies use a MuJoCo-based solver (Todorov et al., 2012), ropes use vertex block descent (VBD) (Chen et al., 2024), cloth uses the Style3D solver in Newton (Newton Contributors, 2025), deformable volumes and liquids use implicit material point methods (MPM) (Stomakhin et al., 2013), and gases use the Mantaflow grid solver (MantaFlow Developers, 2020). The resulting native representations—surface meshes for solids, particles for liquids, and volumetric fields for gases—are converted into standardized visual observations through material-specific rendering: meshes are path-traced directly, liquid particles are reconstructed with SplashSurf (Löschner et al., 2023) for reflection and refraction, and gas fields are rendered as participating media. All rollouts are rendered in Blender (Blender Online Community, 2024) at 1920×1080, 24 fps, and 48 frames, using a fixed 4K equirectangular HDR environment map per sequence. In addition to RGB frames, we provide structured auxiliary channels for reference-based and physics-grounded evaluation: solid and liquid scenes include camera-space normals, metric depth, and object-identity masks, while gas scenes provide continuous opacity channels encoding smoke visibility and foreground occlusion. The simulation defines the reference dynamics, while rendering produces their visual observations and evaluation signals. Solver configurations and output conventions are detailed in Appendices C.3 and C.4.

2.4 PHYSICS-PRESERVING PHOTOREALISTIC TRANSLATION

Synthetic appearance can introduce a visual-domain bias into physical evaluation. Following Hui et al. (2026), we use Cosmos-Transfer2.5-2B (NVIDIA, 2025) to generate appearance variants conditioned on the source RGB, estimated depth, and material prompts. Depth conditioning and explicit prompt constraints aim to retain scene geometry, viewpoint, and motion while varying visual materials. This broadens the appearance distribution available for evaluation and fine-tuning without redefining the simulation-derived physical target. Each source rollout used in translation receives eight variants. Prompt construction and translation settings are given in Appendix C.5.

2.5 BENCHMARK STATISTICS AND QUALITY CONTROL

The benchmark comprises 960 source scenes across twelve scenario families, with 128 scenes per rigid- or soft-body scenario and 32 scenes per liquid or gas scenario (Table 7 in Appendix C.1). Each rollout contains 48 frames at 1920×1080 and 24 fps, spanning 2 s. The appearance-translated set contains 6,272 videos. Quality control combines *pre-compilation validation* of initial configurations with *manual inspection* of complete rollouts. Invalid states, interpenetration, inadequate visibility, incorrect physical evolution, and uninformative randomizations are rejected. Scene-level seed ranges isolate resampling from other instances (Appendix C.6).

3 REFERENCE-BASED PHYSICAL EVALUATION

We measure fidelity to the *observable evolution* of a specified simulation reference, rather than the physical plausibility of alternative futures. Because physical errors manifest differently across materials, our framework combines material-specific geometric observables with their temporal evolution. This separates the aspects of reference disagreement that a single appearance-based metric may conflate.

3.1 EVALUATION FORMULATION

Each sample pairs a reference rollout with a generated continuation conditioned on the same initial image. Let $G_t, P_t \subseteq \Omega$ denote their visible regions at reference frame t. We compare them using a fixed spatial mapping and the declared frame rates, without per-frame registration or temporal warping, preserving position and timing errors. Reference regions come from benchmark annotations; predicted regions are supplied as masks or extracted using SAM 2 (Ravi et al., 2025).

For physical evaluation of generated videos, whole-frame similarity mixes appearance variation with physical disagreement: lighting may change without changing motion, while a localized motion

error may affect few pixels (Motamed et al., 2025; Jain & Wu, 2026). We therefore compare visible geometry and its evolution, retaining SSIM and RGB-frame-difference discrepancies as baselines. Observation and alignment conventions are given in Appendix D.1.

3.2 Material-Specific Observables

Rigid bodies. Following object-centric evaluation (Jain & Wu, 2026; Li et al., 2025), we combine IoU, centroid distance, and foreground Chamfer distance. The distance terms remain informative when masks are disjoint or their centroids coincide despite different shapes. Surface-point trajectories additionally capture motion, including rotation, that silhouettes may miss.

Soft bodies. Correct placement does not imply correct deformation. We combine overlap and boundary agreement with initially normalized width, height, and area curves to characterize the magnitude and timing of projected deformation. Presence and component-count measurements describe visible-region consistency.

Liquids. IoU measures global support, Boundary IoU (Cheng et al., 2021) emphasizes fronts, and the Fractions Skill Score (FSS) (Roberts & Lean, 2008) compares occupancy at multiple spatial scales. To distinguish spatial overlap from actual evolution, we also apply FSS to support changes between successive distinct predicted observations. This temporal FSS penalizes frozen support when the reference changes, while remaining insensitive to motion within unchanged support.

Smoke. We threshold reference opacity and compare the resulting visible support with RGB-derived predicted regions. Coarse occupancy and temporal FSS assess support distribution and change; spread and onset diagnostics characterize plume expansion. These measurements do not recover opacity amplitude or physical density. Detailed definitions are in Appendix D.

3.3 Unified Physical Consistency Score

Defined scores lie in $[0, 100]$, with higher values indicating closer reference agreement; this shared range does not calibrate errors across materials. Components and observation coverage accompany each aggregate. The rigid-body score weights overlap, foreground geometry ,centroid agreement and point motion equally. For soft bodies, $\mathcal{G}$ measures geometry, $\mathcal{D}$ displacement and deformation response, and $\mathcal{I}$ visible-instance consistency. For liquids, $\mathcal{S}$ combines spatial metrics and $\mathcal{T}$ summarizes temporal FSS:

$$S_{\text{soft}} = 100\,\mathcal{G}^{0.50}\mathcal{D}^{0.35}\mathcal{I}^{0.15}, \tag{1}$$

$$S_{\text{liquid}} = 100\sqrt{\mathcal{S}\mathcal{T}}. \tag{2}$$

The geometric means prevent complete failure in one factor from being compensated by another. Weights and error scales remain design choices requiring empirical assessment. The smoke score combines IoU with coarse-scale occupancy and temporal FSS. Aggregation details appear in Appendix D.6.

3.4 Validation against Human Judgment

A development-stage study compares Wan-2.2 TI2V-5B and CogVideoX-1.5-5B I2V on 200 inputs, with 50 each for rigid bodies, soft bodies, liquids and smoke. One evaluator judges physical plausibility rather than reference fidelity. After excluding invalid and tied pairs, soft-body and liquid aggregate scores agree with human preferences at 93.6% and 98.0%, respectively; rigid-body geometric metrics achieve 71.4–81.6%. These results are supporting evidence from one evaluator and one model pair, not a general validation of reference-fidelity ranking. Full results are reported in Appendix E.

4 Benchmarking Video Generation and World Models

4.1 Experimental Setup

Models. We evaluate seven open-weight video generators—Wan2.2-TI2V-5B, Cosmos-Predict2.5-2B, Cosmos3-Edge, HunyuanVideo-1.5, CogVideoX1.5-5B-I2V, Wan2.1-I2V-14B, and

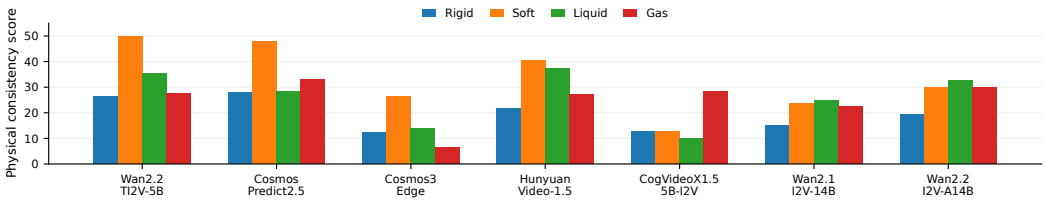


Figure 2: Performance of seven video generation models across four material categories.

Table 2: Physical consistency on rigid-body and soft-body scenarios (unified total score, %). Higher is better. Rigid-body entries use the unified rigid score (rollout IoU, Chamfer, mask-centroid L2 and point ATE); soft-body entries use the soft-body score. Best results are shown in bold and second-best results are underlined.

	Rigid-body				Soft-body			
Model	Falling Objects	Rolling Objects	Dominoes	Avg.	Soft Pad	Cloth (flag)	Rope	Avg.
Wan2.2-TI2V-5B	**42.1**	**52.3**	<u>50.6</u>	**48.3**	<u>76.4</u>	**41.4**	**31.8**	**49.9**
Cosmos-Predict2.5-2B	<u>40.9</u>	<u>51.8</u>	**51.8**	<u>48.2</u>	**81.1**	32.2	30.3	<u>47.9</u>
Cosmos3-Edge	25.8	16.0	39.2	27.0	51.2	6.9	20.5	26.2
HunyuanVideo-1.5	38.9	43.0	49.8	43.9	55.7	<u>33.9</u>	<u>31.6</u>	40.4
CogVideoX1.5-5B-I2V	24.3	24.6	34.9	27.9	19.3	11.3	7.5	12.7
Wan2.1-I2V-14B	27.1	36.7	31.9	31.9	41.9	17.4	11.9	23.7
Wan2.2-I2V-A14B	30.1	48.4	47.2	41.9	57.5	12.3	19.3	29.7

Wan2.2-I2V-A14B—on the same benchmark instances with identical initial images and prompts. Each model uses its official inference implementation without benchmark-specific fine-tuning, retaining its native resolution, frame rate, sampling steps, and guidance scale. Detailed configurations are provided in Appendix I.

4.2 Main Results

Figure 2 shows material-dependent performance across the seven models: Wan2.2-TI2V-5B leads on rigid-body and soft-body scenarios, Cosmos-Predict2.5-2B on gas scenarios, and HunyuanVideo-1.5 on liquids. These differences highlight the importance of material-specific evaluation rather than a single overall ranking. Tables 2–3 report scenario-level physical consistency scores and category averages for seven models across rigid-body, soft-body, and fluid dynamics. Liquids and gases are grouped under *fluid dynamics* for presentation but remain distinct in the benchmark taxonomy. Higher scores indicate closer agreement with the corresponding simulation-derived oracle rollouts.

Rigid-body dynamics. Wan2.2-TI2V-5B obtains the best rigid-body average (48.3), leading on *falling objects* (42.1) and *rolling objects* (52.3), while Cosmos-Predict2.5-2B is a close second overall (48.2) and takes *dominoes* (51.8; Table 2). *Falling objects* remains the hardest scenario (mean 32.7 over the seven models, best score 42.1), highlighting difficulties in maintaining consistent trajectories through collisions and subsequent motion. Cosmos3-Edge and CogVideoX1.5-5B-I2V trail with mean scores of 27.0 and 27.9, respectively.

Soft-body dynamics. Wan2.2-TI2V-5B achieves the highest mean score (49.9; Table 2), but no model performs consistently well across all three scenarios. The soft-pad scenario exhibits the largest inter-model gap (81.1 versus 19.3), highlighting substantial variation in predicting deformation and elastic response.

Fluid dynamics. Performance varies substantially across liquid and gas scenarios (Table 3). HunyuanVideo-1.5 achieves the highest overall mean (32.3), while Cosmos-Predict2.5-2B leads on gas scenarios, scoring 48.6 on smoke-stack and 43.1 on flame-heating. Despite visually plausible appearance, fluid predictions can accumulate spatial errors over time, increasingly deviating from the oracle rollout.

Table 3: Physical consistency on fluid-dynamics scenarios, including liquid and gas dynamics. Higher is better. Best results are shown in bold and second-best results are underlined.

	Liquid			Gas			
Model	Water Tank	Draining	U-Connector	Smoke Stack	Impact Smoke	Flame Heating	Avg.
Wan2.2-TI2V-5B	**30.7**	<u>41.3</u>	<u>34.5</u>	39.4	<u>10.1</u>	32.6	<u>31.4</u>
Cosmos-Predict2.5-2B	27.5	26.5	30.9	**48.6**	7.9	**43.1**	30.8
Cosmos3-Edge	15.6	14.8	11.2	10.2	1.6	7.1	10.1
HunyuanVideo-1.5	<u>28.8</u>	38.8	**44.5**	42.2	9.7	30.0	**32.3**
CogVideoX1.5-5B-I2V	7.3	10.8	12.0	40.8	**12.4**	32.3	19.3
Wan2.1-I2V-14B	19.0	30.2	25.1	37.5	4.5	25.2	23.6
Wan2.2-I2V-A14B	24.3	**41.8**	32.0	<u>45.2</u>	9.4	<u>34.6</u>	31.2

Takeaway. Overall, the results highlight a clear distinction between visual plausibility and physical consistency. Although all evaluated models can generate visually convincing motion, their agreement with the physically grounded oracle varies substantially across scenarios and material classes. The strong variation in model rankings across rigid, soft, and fluid dynamics further indicates that physical capability is multi-dimensional: success on one class of physical phenomena does not imply robust physical prediction in others.

Cross-material comparison. The reported mean over rigid- and soft-body scenarios is highest for Wan2.2-TI2V-5B (49.1), closely followed by Cosmos-Predict2.5-2B (48.0), and lowest for CogVideoX1.5-5B-I2V (20.3). These are descriptive averages: the material-specific metrics are not calibrated to a common physical-error scale. The more informative finding is the change in within-category rankings, which shows that strong reference agreement in one physical regime does not imply the same relative performance in another. VidPhysOracle therefore characterizes physical prediction through material-specific performance rather than a single overall ordering.

4.3 Comparison with Existing Physical Evaluators

Table 1 compares evaluation setup, material coverage, and metric design. VBench and VBench-2.0 assess general video quality, with the latter additionally evaluating physical consistency (Huang et al., 2023; Zheng et al., 2025). VideoPhy and PhyGenBench target physical interactions without paired reference videos (Bansal et al., 2024; Meng et al., 2024). In the reported setups, Physion-Eval and Physics-IQ use reference videos (Zhang et al., 2026; Motamed et al., 2026), while RigidBench additionally provides explicit rigid-body measurements (Jain & Wu, 2026).

VidPhysOracle combines simulation-derived references with material-specific metrics across rigid bodies, soft bodies, liquids, and gases, enabling broader analysis of observable physical evolution. Entries marked "/" indicate unspecified coverage under our taxonomy, not exclusion of the corresponding phenomena.

4.4 Does Photorealistic Translation Reduce the Visual Domain Gap?

Qualitative comparisons suggest that appearance transfer reduces simulation-specific rendering artifacts. Liquid scenes exhibit smoother surfaces and more natural transparency, reflections, and refractions, while rigid-body scenes gain realistic materials, lighting, and shadows.

Depth conditioning and composition-preserving prompts encourage consistency in scene layout, viewpoint, and motion, as illustrated by the rope example (Figure 7). Paired source and translated videos therefore support testing whether physical predictions remain robust to appearance changes. Performance degradation after translation may indicate reliance on simulation-specific visual cues, although transfer-induced artifacts remain a potential confound. Appearance transfer thus mitigates the visual component of the sim-to-real gap, but neither corrects underlying simulation errors nor guarantees exact structural and temporal preservation. Full details are provided in Appendix C.5

Table 4: Resolution ablation of Mask IoU. All lower-resolution videos are generated from the same 1080p reference. Δ denotes the absolute change from 1080p to 288p.

Category	Mask source	1080p	720p	576p	432p	360p	288p	Δ
Rigid	GT resampling	1.0000	0.9874	0.9843	0.9776	0.9762	0.9673	-0.0327
Soft	GT resampling	1.0000	0.9835	0.9790	0.9720	0.9690	0.9593	-0.0407
Fluid	GT resampling	1.0000	0.9945	0.9928	0.9905	0.9895	0.9856	-0.0144
Rigid	SAM2	0.9749	0.9637	0.9563	0.9463	0.9451	0.9379	-0.0370
Soft	SAM2	0.9703	0.9624	0.9528	0.9431	0.9399	0.9332	-0.0371
Fluid	SAM2	0.9533	0.9532	0.9465	0.9339	0.9308	0.9266	-0.0267

4.5 Ablation Studies

To assess resolution sensitivity independently of physical prediction errors, we vary spatial resolution while keeping scene content and motion fixed.

Experimental setup. We select three examples from each of eight scenarios, yielding 24 videos and 35 annotated objects: nine rigid-body, nine soft-body, and six fluid videos. Each 1920×1080 reference is downsampled to 1280×720, 1024×576, 768×432, 640×360, and 512×288, spanning a $14.1\times$ range in pixel count.

We compare two settings. *GT-resampling* uses area-downsampled ground-truth masks thresholded at 0.5, isolating rasterization effects. *SAM2* additionally captures segmentation error by running SAM2.1 Large independently at each resolution, prompted with the corresponding ground-truth first-frame mask. All masks are mapped to a common 1920×1080 evaluation canvas and scored over frames 1–47, with frame rate, temporal alignment, and metric parameters fixed.

Results. Reducing resolution from 1080p to 288p decreases GT-resampling IoU by 3.27, 4.07, and 1.44 percentage points for rigid-body, soft-body, and fluid scenes, respectively (Table 4). At 288p, IoUs remain 0.967, 0.959, and 0.986, while centroid errors remain below 0.05% of image height across categories. The SAM2 pipeline introduces additional segmentation error but exhibits similarly modest IoU reductions of 3.70, 3.71, and 2.67 percentage points. These results indicate limited resolution sensitivity for the evaluated spatial metrics under fixed motion and common-canvas alignment.

Resolution sensitivity of physical metrics. At 288p, GT-resampling retains soft-body and fluid scores of $98.49/100$ and $93.60/100$, respectively. Across resolutions, fluid occupancy FSS exceeds 0.9998 at $0.02H$ and 0.9999 at $0.10H$, while projected-area curve MAE remains below 2×10^{-4}. By comparison, rigid-body SSIM decreases from 0.996 to 0.923. These results support limited resolution sensitivity of the evaluated physical observables, although changes in differently scaled metrics are not directly comparable.

Scale-aware boundary tolerance. On six fluid scenes, increasing the boundary tolerance improves robustness to rasterization (Table 5). At 288p, a fixed 5-pixel band yields Boundary IoU of 0.729, compared with 0.962 using $2\%D$, where D is the evaluation-canvas diagonal. This tolerance retains sensitivity to localization errors: translating the original 1080p masks by 11 pixels reduces Boundary IoU to 0.790, a score drop $5.53\times$ larger than that caused by downsampling. These results support scale-aware tolerances for balancing rasterization robustness with sensitivity to spatial errors.

Discussion. The ablation supports approximate resolution independence of the principal spatial and scale-normalized physical metrics over the tested 288p–1080p range. However, this property does not hold equally for every diagnostic. Narrow-band boundary metrics remain sensitive to rasterization, and event timing may vary when a threshold crossing occurs near a frame boundary; for example, the soft-body event-time MAE reaches $0.253\,\mathrm{s}$ at 288p under GT resampling. We therefore do not claim strict resolution invariance for every metric. Instead, the results suggest

Table 5: Boundary IoU under different boundary tolerances on fluid scenes. D denotes the diagonal length of the common 1920×1080 evaluation canvas. The last column introduces a genuine localization error by translating the 1080p masks by 11 pixels.

Boundary tolerance	1080p	720p	576p	432p	360p	288p	11-px shift
5 px ($0.005H$)	1.0000	0.8907	0.8557	0.8116	0.7940	0.7289	0.2013
11 px ($0.01H$)	1.0000	0.9462	0.9287	0.9051	0.8958	0.8606	0.3704
$1\%D$	1.0000	0.9716	0.9626	0.9499	0.9448	0.9253	0.6223
$2\%D$	1.0000	0.9857	0.9812	0.9745	0.9721	0.9620	0.7899

Table 6: **Effect of LoRA adaptation on Wan 2.2 TI2V 5B.** Comparison between the base and LoRA-adapted models on the soft-pad evaluation set.

Physical consistency scores				**Rollout and geometric metrics**			
Metric	Base	LoRA	Δ	Metric	Base	LoRA	Δ
Soft-body total ↑	34.05	**89.76**	+55.71	Rollout IoU ↑	0.4118	**0.8822**	+0.4704
Geometry ↑	0.3679	**0.8664**	+0.4985	Boundary Chamfer / H ↓	0.3398	**0.0066**	-0.3332
Trajectory ↑	0.2947	**0.9419**	+0.6472	Boundary HD95 / H ↓	0.4340	**0.0340**	-0.4000
Deformation ↑	0.3564	**0.8686**	+0.5122	Centroid distance / H ↓	0.3816	**0.0060**	-0.3755
Dynamics ↑	0.2967	**0.9042**	+0.6074	Object missing rate ↓	0.1335	**0.0000**	-0.1335
Integrity ↑	0.8587	**0.9951**	+0.1364				

that videos generated at different resolutions can be compared reliably when spatial predictions are mapped to a common coordinate system and pixel-scale tolerances are normalized by image size.

5 Fine-Tuning with VidPhysOracle

To assess VidPhysOracle as a training resource, we fine-tune Wan 2.2 TI2V 5B with LoRA on two soft-body scenarios. The training loss decreases overall (Figure 3). We compare the adapted model (**LoRA**) with the original (**Base**) on 128 samples from `falling_objects_on_soft_pad` using our soft-body metrics. Table 6 and Figure 14 present quantitative and qualitative comparisons, respectively. Full training and evaluation details are provided in Appendix H.

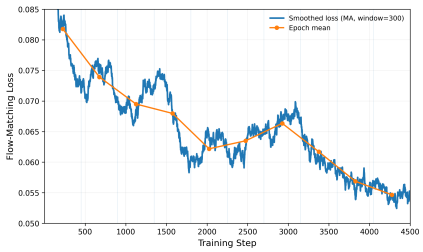


Figure 3: Smoothed flow-matching loss during LoRA fine-tuning.

6 Conclusion

We presented VidPhysOracle, a simulation-grounded benchmark and video dataset for evaluating physical prediction across rigid bodies, soft bodies, liquids, gases, and their interactions. Controlled oracle rollouts shift evaluation from visual plausibility to fidelity to specified physical processes. Our material-aware metrics combine object geometry, surface-point motion, deformation response, and visible-support evolution to diagnose complementary prediction errors. Temporal support comparison further penalizes frozen predictions when reference support changes, even under substantial spatial overlap. Simulation-derived annotations and photorealistic appearance variants support both benchmarking and physics-aware fine-tuning. Together, these contributions connect controlled reference generation, interpretable evaluation, and model adaptation to advance physically grounded video world modeling.

Our metrics assess two-dimensional observations rather than complete physical states. Future work could reconstruct velocity, density, and deformation fields for equation-based validation while accounting for reconstruction uncertainty. Appearance transfer remains less reliable for smoke and some liquid scenarios, limiting their photorealistic coverage. Improving transfer fidelity and expanding these subsets would support broader evaluation and physics-aware training.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery and synthesis, feedback on research methodology and experimental design, software and code development, organization and interpretation of experimental results, manuscript structuring, translation, language editing and synthetic data generation. All AI-assisted code, analyses, citations, claims, figures, and text were independently reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all material produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies the physical consistency of video generative models and develops evaluation protocols for measuring whether generated videos preserve the dynamics and physical properties implied by a given initial state. Our experiments primarily use procedurally generated simulation data and outputs from publicly available or research-accessible video generation models. The proposed metrics evaluate physical behavior and visual consistency and do not rely on personally identifiable information or sensitive attributes.

No human-subject data are used for model training or metric construction. When human evaluation is employed to compare the physical plausibility of generated videos, participants are only asked to judge predefined video pairs and are not required to provide personal or sensitive information. The study does not involve medical, biometric, surveillance, or other high-risk applications.

REPRODUCIBILITY STATEMENT

The main paper and appendices describe benchmark construction, evaluation protocols, metric definitions, model configurations, and fine-tuning settings. We report the data splits, input conditions, and inference settings used in our experiments. The supplementary material provides some video samples of our dataset. And we promise to release code for metric evaluation.

REFERENCES

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

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. doi: 10.48550/arXiv.2406.03520. URL https://arxiv.org/abs/2406.03520.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2503.06800.

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. URL https://arxiv.org/abs/2311.15127.

Blender Online Community. Blender: A 3d modelling and rendering software. Software, 2024. URL https://www.blender.org/. Version 4.4.1; accessed 2026-09-23.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and

Tim Rocktäschel. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. doi: 10.48550/arXiv.2402.15391. URL https://arxiv.org/abs/2402.15391.

Anka He Chen, Ziheng Liu, Yin Yang, and Cem Yuksel. Vertex block descent. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10. ACM, 2024. doi: 10.1145/3610548.3618220.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://arxiv.org/abs/2103.16562.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. doi: 10.48550/arXiv.2204.03458. URL https://arxiv.org/abs/2204.03458.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023. doi: 10.48550/arXiv.2311.17982. URL https://arxiv.org/abs/2311.17982.

Chenyu Hui, Xiaodi Huang, Siyu Xu, Yunke Wang, Shan You, Fei Wang, Tao Huang, and Chang Xu. Seeing realism from simulation: Efficient video transfer for vision-language-action data augmentation, 2026. URL https://arxiv.org/abs/2605.02757.

Swarnim Jain and Shangzhe Wu. RigidBench: Evaluating rigid-body physics in video generation models. *arXiv preprint arXiv:2608.15555*, 2026. doi: 10.48550/arXiv.2608.15555. URL https://arxiv.org/abs/2608.15555.

Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. The affine particle-in-cell method. *ACM Transactions on Graphics*, 34(4):1–10, 2015. doi: 10.1145/2766996.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. doi: 10.48550/arXiv.2411.02385. URL https://arxiv.org/abs/2411.02385.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025. URL https://arxiv.org/abs/2503.09595.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. *arXiv preprint arXiv:2409.18964*, 2024. doi: 10.48550/arXiv.2409.18964. URL https://arxiv.org/abs/2409.18964.

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. doi: 10.2312/vmv.20231245.

MantaFlow Developers. Mantaflow: A framework for fluid simulation. Open-source software repository, 2020. URL https://github.com/thunil/mantaflow. Accessed 2026-09-23.

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

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. URL https://arxiv.org/abs/2501.09038.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. URL https://physics-iq.github.io/.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. PhyCo: Learning controllable physical priors for generative motion. *arXiv preprint arXiv:2604.28169*, 2026. doi: 10.48550/arXiv.2604.28169. URL https://arxiv.org/abs/2604.28169.

Newton Contributors. Newton: A gpu-accelerated physics simulation engine. Software repository, 2025. URL https://github.com/newton-physics/newton. Accessed 2026-09-23.

NVIDIA. Cosmos-Transfer2.5-2B. Hugging Face model card, 2025. URL https://huggingface.co/nvidia/Cosmos-Transfer2.5-2B. Accessed: 2026-09-17.

Tim Rädsch, Yuki M. Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-IQ Verified. *arXiv preprint arXiv:2606.18943*, 2026. doi: 10.48550/arXiv.2606.18943. URL https://arxiv.org/abs/2606.18943.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Nigel M. Roberts and Humphrey W. Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, 2008. doi: 10.1175/2007MWR2123.1.

Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics*, 32(4):1–10, 2013. doi: 10.1145/2461912.2461948.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(29), 2015. doi: 10.1186/s12880-015-0068-x.

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

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Shuai Wang, Yaxin Feng, Xuekun Jiang, Shihan Tian, Ningyu Yan, Xing Shen, Chaoyang Lyu, Hui Wang, Yunsong Zhou, Hanqing Wang, Jiangmiao Pang, Yang Xiang, Xing Gao, Chunhua Shen, and Weinan Zhang. GAUGE: A measurement-grounded benchmark for physical fidelity in simulation engines and video world models. *arXiv preprint arXiv:2608.05948*, 2026. doi: 10.48550/arXiv.2608.05948. URL https://arxiv.org/abs/2608.05948.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025. doi: 10.48550/arXiv.2511.18870. URL https://arxiv.org/abs/2511.18870.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. doi: 10.48550/arXiv.2310.06114. URL https://arxiv.org/abs/2310.06114.

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. *arXiv preprint arXiv:2408.06072*, 2024. doi: 10.48550/arXiv.2408.06072. URL https://arxiv.org/abs/2408.06072.

Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918v2*, 2025. URL https://arxiv.org/abs/2504.02918v2.

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

Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis. *arXiv preprint arXiv:2503.20822*, 2025. doi: 10.48550/arXiv.2503.20822. URL https://arxiv.org/abs/2503.20822.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. doi: 10.48550/arXiv.2503.21755. URL https://arxiv.org/abs/2503.21755.

Siyuan Zhou, Hejun Wang, Hu Cheng, Jinxi Li, Dongsheng Wang, Junwei Jiang, Yixiao Jin, Jiayue Huang, Shiwei Mao, Shangjia Liu, Yafei Yang, Hongkang Song, Shenxing Wei, Zihui Zhang, Peng Huang, Shijie Liu, Zhengli Hao, Hao Li, Yitian Li, Wenqi Zhou, Zhihan Zhao, Zongqi He, Hongtao Wen, Shouwang Huang, Peng Yun, Bowen Cheng, Pok Kazaf Fu, Wai Kit Lai, Jiahao Chen, Kaiyuan Wang, Zhixuan Sun, Ziqi Li, Haochen Hu, Di Zhang, Chun Ho Yuen, Bing Wang, Zhihua Wang, Chuhang Zou, and Bo Yang. PhysInOne: Visual physics learning and reasoning in one suite. *arXiv preprint arXiv:2604.09415*, 2026. doi: 10.48550/arXiv.2604.09415. URL https://arxiv.org/abs/2604.09415.

A Related Work

Our work connects video world modeling, physical evaluation, and simulation-grounded training.

A.1 Video Generation and World Models

Video diffusion models improve visual quality and temporal coherence through scalable architectures, latent representations, and data curation (Ho et al., 2022; Blattmann et al., 2023; Yang et al., 2024; Team Wan et al., 2025; Wu et al., 2025). UniSim, Genie, and Cosmos extend video generation toward interactive simulation and physical world modeling (Yang et al., 2023; Bruce et al., 2024; Ali et al., 2025). However, visual realism does not ensure generalization of physical laws or accurate dynamics (Kang et al., 2024; Motamed et al., 2026), motivating evaluation beyond appearance.

A.2 Physical Evaluation of Generated Videos

VBench evaluates visual and temporal quality (Huang et al., 2023), while VBench-2.0 incorporates physics and commonsense (Zheng et al., 2025). Physics-oriented methods primarily rely on explicit constraints, human or learned judgments, or reference trajectories; these approaches may share learned perception components.

Rule-based and physics-informed metrics. Morpheus assesses physical invariants and governing equations using recovered trajectories (Zhang et al., 2025). Such checks are interpretable but depend on observable states and suitable physical models, with deformation and flow requiring different representations. Moreover, satisfying selected constraints does not establish fidelity to specified initial conditions and material parameters. VidPhysOracle complements these checks with simulation-referenced evaluation.

Judge-based physical evaluators. VideoPhy and VideoPhy-2 assess material interactions and action-centric physical commonsense (Bansal et al., 2024; 2026), while PhyGenEval combines vision-language and language models for hierarchical assessment (Meng et al., 2024). Physion-Eval provides localized expert reasoning and identifies limitations of automated critics (Zhang et al., 2026). These approaches offer broad coverage but depend on recognizing physical violations rather than directly measuring deviation from a specified evolution.

Reference-based and simulation-grounded evaluation. Physics-IQ uses real-world experimental references (Motamed et al., 2026), and Physics-IQ Verified examines reference quality, prompts, and aggregation (Rädsch et al., 2026). RigidBench supplies simulated rigid-body rollouts, state annotations, and training videos (Jain & Wu, 2026). GAUGE covers rigid and deformable systems, although its video-model evaluation uses a rigid-body subset (Wang et al., 2026). VidPhysOracle combines cross-material reference generation with material-specific measurements of observable dynamics. Its objective is fidelity to the specified simulation, not the validity of every possible continuation.

A.3 Physics-Grounded Generation and Training Data

PhysGen grounds generation in rigid-body simulation, rendering, and diffusion refinement (Liu et al., 2024). For model adaptation, SimDrop mitigates synthetic appearance artifacts (Zhao et al., 2025), PhysInOne provides physically annotated videos (Zhou et al., 2026), and PhyCo combines simulation-supervised fine-tuning, physical-property conditioning, and vision-language reward optimization (Narayanan et al., 2026). RigidBench also studies simulation-grounded fine-tuning (Jain & Wu, 2026).

VidPhysOracle connects evaluation and adaptation through paired initial observations and simulated futures. Cosmos-based appearance transfer broadens visual diversity while simulation defines the reference dynamics (NVIDIA, 2025; Ali et al., 2025). Preserving geometry and motion remains a data-quality requirement, not an automatic guarantee of appearance transfer.

Table 7: **Scenario suite and benchmark statistics.** The table lists the twelve scenario families, their source scenes, appearance- translated videos, physical targets, and annotated entities. All scenes provide RGB and object-identity manifests. Solid and liquid scenes additionally provide normals, metric depth, and identity masks; gas scenes provide continuous opacity channels.

Scenario	Scenes	Trans.	Physical targets	Entities
Rigid bodies				
Falling_objects	128	1024	gravity, collision, momentum conservation, resting stability	2–3 bodies
Rolling_objects	128	1024	gravity, friction, collision, momentum conservation	2 bodies, ramp
Dominoes	128	1024	collision, momentum conservation, friction, resting stability	2–4 boxes
Soft bodies				
Falling_objects_on_soft_pad	128	1024	soft–rigid interaction, elastic recovery, damping and energy dissipation	pad, 2 bodies
Gravity_draping_flag	128	896	draping, wrinkling, bending, inextensibility	cloth, pole
Gravity_tightening_soft_rope	128	1024	bending, inextensibility, damping and energy dissipation	rope, load, pole
Liquids				
Rigid_objects_in_water_tank	32	–	buoyancy, liquid–solid interaction, free-surface evolution	liquid, 2 bodies
Draining_tank	32	256	free-surface evolution, advection, mass conservation, pressure	liquid, stand
U_connector	32	–	pressure, mass conservation, advection, liquid–solid interaction	liquid
Gases				
Smoke_stack	32	–	thermal convection, buoyancy, advection, dispersion and mixing	smoke, container
Impact_smoke_dispersion	32	–	vortex dynamics, dispersion and mixing, gas–solid interaction	smoke, body
Flame_heating	32	–	combustion, thermal convection, buoyancy, gas–solid interaction	smoke, flame, opaque objects
Total	**960**	**6272**		

B Appendix

C Benchmark Construction Details

C.1 Scenario Suite and Coverage

Table 7 lists the twelve scenario families, their appearance- translated videos, physical targets, and annotated entities. Three families represent each material group; coupled interactions are embedded within them rather than counted as additional scenarios. Soft-body cases span 1D ropes, 2D cloth, and 3D volumes. Figure 4 shows representative frames.

Comparison with existing benchmarks. Table 1 summarizes the selected benchmarks under the coverage categories used in this work. The table distinguishes reference-video availability and rule-based metrics from material coverage.

C.2 Scenario Diversification and DSL

Scenario families are designed around controlled initialization, a small set of identifiable physical behaviours, and independent appearance variation. To support these requirements, we use a three-

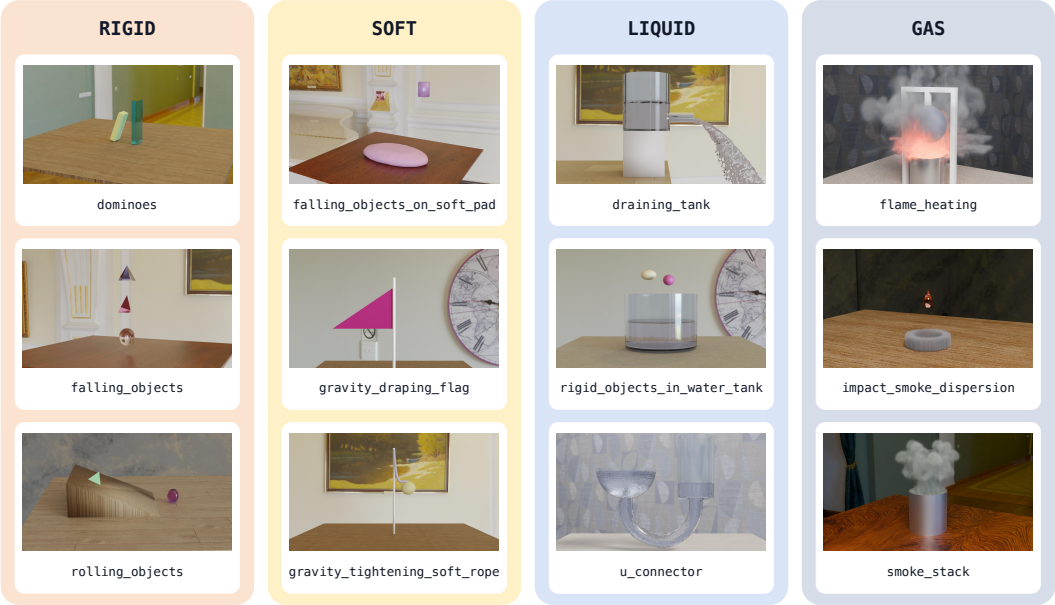


Figure 4: **Scenario examples.** Representative frames from the twelve families, grouped into rigid-body, soft-body, liquid, and gas dynamics; each column contains three scenarios.

layer Domain-Specific Language (DSL) that separates reusable object templates, scenario composition, and global rendering conditions (Figure 5). Scenes start from rest without prescribed velocities or keyframed motion, while geometry, layout, and appearance are randomized independently at their respective DSL layers. Textures, colours, roughness, metallicness, and environment maps can therefore vary across scenes without altering the underlying physical parameters, enabling diverse visual realizations of the same physical setup.

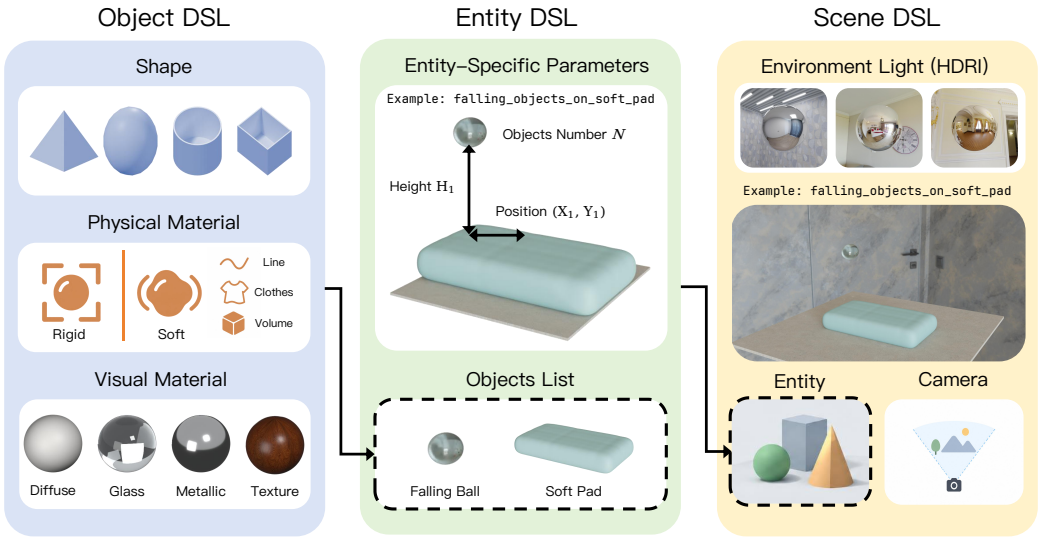


Figure 5: **DSL construction.** The three-layer DSL separates reusable object templates, scenario composition, and global rendering conditions.

Object DSL. Reusable templates combine shape, physical material, and visual material, parameterizing object-level geometry and appearance.

Table 8: **Reference solver configurations.** Rates are simulation rates, not video frame rates. Iteration counts retain the source's per-step convention; substeps are listed separately.

System	Backend	Configuration
Rigid bodies	Newton's MuJoCo-based solver (Todorov et al., 2012)	240 Hz; 200 solver iterations; up to 50 CCD iterations per step
Ropes	Newton's vertex block descent (VBD) (Chen et al., 2024)	32 substeps; 25 iterations per step
Cloth	Style3D cloth solver in Newton (Newton Contributors, 2025)	25 substeps; 10 Newton and 20 linear iterations per step
Soft volumes	Newton's implicit material point method (MPM) (Stomakhin et al., 2013), with affine particle-in-cell (APIC) transfer (Jiang et al., 2015)	240 Hz; 200 iterations per step
Liquids	Implicit MPM	120 Hz; particle counts reaching millions per scene
Gases	Mantaflow grid solver (MantaFlow Developers, 2020)	Up to 2.3 million cells

Entity DSL. Each family specifies an object list and spatial arrangement, together with family-specific randomization parameters such as object count, position, and initial height.

Scene DSL. The global environment is selected from seven indoor maps and held fixed as both illumination and background throughout the rollout. An Entity-specific camera preset keeps the relevant objects and their expected motion envelope within view. Solver settings remain fixed within each family; geometry, layout, and appearance vary only through their designated parameters.

C.3 Reference Simulation Configurations

Table 8 lists the material-specific backends and numerical settings. These configurations prioritize stable reference rollouts rather than real-time throughput. Rigid-body simulation additionally uses full inertia tensors and explicit penetration correction.

C.4 Rendering and Evaluation Outputs

Material-specific rendering. Rigid and soft bodies are represented as surface meshes, liquids as particles, and gases as volumetric fields. Solid meshes are path-traced directly. Liquid particles are reconstructed into surfaces using SplashSurf (Löschner et al., 2023) for reflection and refraction. Gas fields are rendered as participating media, integrating scattering and absorption along camera rays. This rendering stage converts the solver outputs into visual observations without defining the simulated dynamics.

Rendering settings. Blender (Blender Online Community, 2024) produces 48 frames at 1920×1080 and 24 fps per source rollout, with denoising and up to 96 samples per pixel. A fixed 4K equirectangular high-dynamic-range environment map supplies both lighting and background within each sequence; visual attributes may vary across sequences.

Auxiliary channels. Each scene includes numbered RGB frames and an object-identity manifest. For solid and liquid scenes, we export camera-space normals and metric depth in 16-bit OpenEXR, together with object-identity masks. Liquid masks separate the reconstructed surface from surrounding solids, including glass containers. For gases, continuous opacity channels replace binary masks: the first channel records smoke opacity, while object channels are attenuated by foreground smoke transmittance. These channels describe visible support and occlusion; they do not constitute a recovered three-dimensional density field.

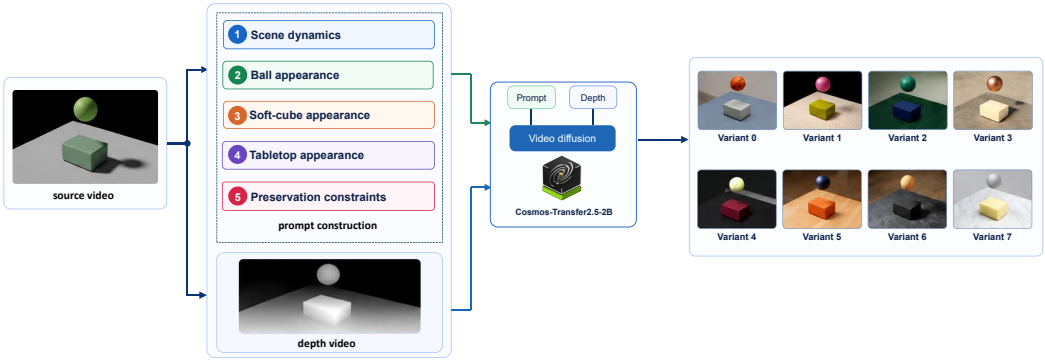


Figure 6: **Photorealistic translation pipeline**. *Left*: source simulation video. *Middle*: metadata-derived appearance prompt specifying materials and structure-preservation constraints. *Right*: the transferred video, intended to vary appearance while retaining source geometry and motion.

C.5 Photorealistic Translation Details

Conditioning. Following Hui et al. (2026), we apply Cosmos-Transfer2.5-2B (NVIDIA, 2025) to source simulation videos using depth control and appearance prompts. Depth sequences are estimated from source RGB using the ViT-S variant of Video Depth Anything (Chen et al., 2025), rather than taken from simulator-exported depth. Simulator metadata is used separately to construct the prompts. The standard pipeline uses no edited first-frame or additional image condition.

For the draping-flag scenario, directly specifying and modifying patterns on the flag fabric in the simulator is cumbersome. We therefore use appearance transfer to introduce flag patterns: FLUX.2-klein-4B (Labs, 2025) edits the flag appearance in the first frame, which is then provided as an additional image condition to Cosmos-Transfer2.5-2B. Depth is estimated from the original simulation video to guide the cloth geometry and motion. Standard inference uses 36 sampling steps, guidance 3, and depth-control weight 1.0. Outputs contain 48 frames at 24 fps with a resolution of 1280×704.

Appearance prompts. For each source video, we construct appearance prompts from scene metadata, combining a photorealistic rendering instruction, scenario-specific dynamics, and appearance specifications for the relevant objects and support surfaces. Where applicable, opaque objects are assigned varied materials, while transparent objects are instructed to preserve their glass appearance. Fluid prompts additionally constrain liquid appearance and vessel geometry. Prompts request preservation of object count, geometry, layout, viewpoint, and motion. Appearance choices are generated deterministically from source and variant identifiers under a recorded prompt policy; prompts and generation seeds are saved for reproducibility. These conditions encourage structural and temporal preservation but do not guarantee it.

Translated-set size. We translate 800 of the 960 source videos: all 128 videos from each of the three rigid-body and three soft-body scenarios, together with 32 videos from the draining-tank scenario. For the draping-flag scenario, quality checks on the edited first frames leave 896 translated videos across its 128 source videos. Generating eight appearance variants per source video yields 6,272 translated videos in total.

C.6 Quality Control and Resampling

Pre-compilation validation. Initial configurations are checked for interpenetration, out-of-range physical or visual parameters, scenario-specific constraint violations, and inadequate camera framing. Rejected samples are redrawn from scene-specific seed ranges, so rejection in one instance does not alter other scenes' randomization.

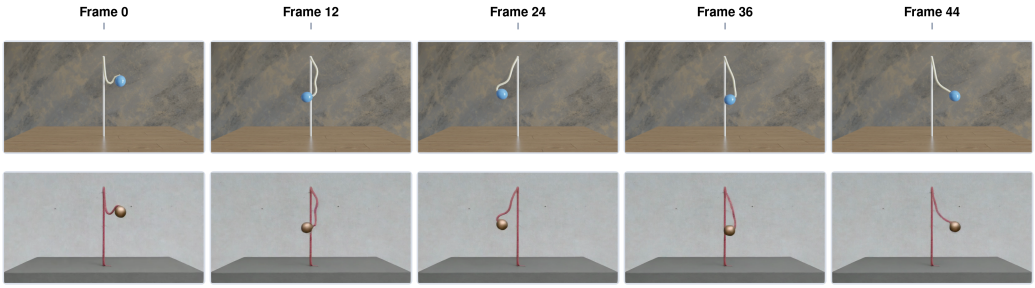


Figure 7: **Comparison between the source simulation and the appearance-transferred video at matched frames.** *Upper:* Simulation frames. *Lower:* Corresponding frames after appearance transfer. The transfer improves visual realism while preserving the rope deformation, object trajectory, and temporal evolution of the original physical event.

Manual rollout inspection. Complete sequences are reviewed for interpenetration, incomplete visibility, incorrect physical evolution, and degenerate initial randomizations that fail to expose the intended behaviour. These checks screen for invalid or uninformative references; they are not presented as a quantitative convergence or simulator-accuracy study.

D Physical Metric Details

D.1 Observation and Alignment Protocol

Let Ω have height H and width W. A single spatial mapping and the declared frame rates place predictions on the reference clock; no per-frame registration or time warping is fitted. Rollout-level mask measurements exclude the conditioning frame, and image-space distances are normalized by H. Reference masks come from scene annotations, with opacity thresholding for smoke. SAM 2 (Ravi et al., 2025), when used, is initialized with the reference frame-0 region. Segmentation and tracking introduce observation errors, so conditional measurements are interpreted together with the observations on which they are defined and their coverage.

D.2 Rigid-Body Measurements

For nonempty regions, overlap and centroid displacement are

$$J_t = \frac{|P_t \cap G_t|}{|P_t \cup G_t|}, \tag{3}$$

$$D_{\mathrm{ctr}}(t) = \frac{\|\mathbf{c}(P_t) - \mathbf{c}(G_t)\|_2}{H}, \tag{4}$$

where $\mathbf{c}(M)$ is the region centroid. Foreground Chamfer distance $D_{\mathrm{CD}}^{\mathrm{fg}}$ compares foreground-pixel distributions, including shape differences that leave the centroid unchanged. These measurements describe visible geometry, not its unique physical cause; orientation and occlusion also affect the masks.

Surface-point motion. Indexed meshes retain surface correspondences across frames. We project these points through the reference trajectory and compare them with tracks from the generated video, capturing motion that region overlap alone may miss.

D.3 Soft-Body Measurements

Boundary Chamfer $D_{\mathrm{CD}}^{\partial}$ measures typical contour displacement; symmetric 95th-percentile distance D_{95}^{∂} captures larger deviations without using the single worst pixel (Taha & Hanbury, 2015). Boundary F1 complements these distances.

For deformation response, bounding-box height, width, and mask area are each normalized by their own initial value in the reference or prediction. Their mean absolute discrepancy defines response error E. Because normalization can conceal constant scale differences, E is interpreted with absolute overlap. These projected measurements depend on pose and occlusion and do not estimate material strain. Presence and component-count errors characterize visible-region consistency, not material topology.

D.4 Liquid Occupancy and Temporal Support

Boundary IoU B_τ (Cheng et al., 2021) emphasizes front alignment. For scale-aware occupancy comparison, FSS (Roberts & Lean, 2008) is

$$\mathrm{FSS}_s(P, G) = 1 - \frac{\sum_x [f_s(P, x) - f_s(G, x)]^2}{\sum_x [f_s(P, x)^2 + f_s(G, x)^2]}, \tag{5}$$

where $f_s(M, x)$ is the foreground fraction in a square window of width approximately sH centered at x. We use $s \in \{0.02, 0.10\}$; larger windows tolerate more local displacement but resolve less detail.

Let $a(t)$ denote the preceding distinct predicted observation on the reference clock. Temporal FSS applies Equation 5 to changed support:

$$\Delta\mathrm{FSS}_s(t) = \mathrm{FSS}_s\big(P_t \triangle P_{a(t)}, G_t \triangle G_{a(t)}\big), \tag{6}$$

where $\triangle$ denotes symmetric difference. A frozen prediction yields zero whenever reference support changes over the interval. This comparison discards the direction of support change and motion within unchanged support.

D.5 Smoke Support and Expansion

Reference support is derived from rendered two-dimensional opacity:

$$G_t = \{x : \alpha_t(x) > \theta\}. \tag{7}$$

The default is $\theta = 0.25$, with 0.10 and 0.50 as sensitivity settings. Coarse occupancy and temporal FSS compare visible support; spread and expansion-onset diagnostics describe its extent and timing. Narrow boundary bands and fine temporal windows are retained as sensitivity diagnostics. Since predictions provide binary support, these measurements estimate neither opacity amplitude nor physical density.

D.6 Score Construction

Distance and response errors are mapped to bounded agreement values before aggregation. The rigid-body score equally weights overlap, foreground geometry, centroid agreement, and surface-point motion, incorporating mask and track coverage. It requires both reference correspondences and predicted tracks.

For Equation 1, $\mathcal{G}$ combines IoU and boundary F1; $\mathcal{D}$ combines centroid and relative-response agreement; and $\mathcal{I}$ measures visible presence and component-count agreement. The exponents and error scales are design choices requiring empirical assessment; $\mathcal{I}$ does not measure material topology.

For Equation 2, $\mathcal{S}$ averages IoU, Boundary IoU, and occupancy FSS after averaging scales within each family. $\mathcal{T}$ averages temporal FSS over its two scales. A zero temporal term forces a zero aggregate. The liquid score applies to references with visible, changing support and predictions covering the reference duration. Equation 11 defines the smoke score.

D.7 Appearance and Temporal Baselines

SSIM measures frame-level appearance agreement. The temporal baseline measures the discrepancy between consecutive RGB-frame differences in the reference and prediction. Neither isolates material-specific geometry or support.

Table 9: Agreement with human preferences for rigid-body and soft-body cases. Tied and invalid pairs are excluded.

Metric	Agreement (%)
Rigid bodies	
IoU	81.6
Centroid distance	71.4
Foreground Chamfer	75.5
Soft bodies	
IoU	91.5
Boundary Chamfer	95.7
Boundary HD95	91.5
Boundary F1	91.5
Centroid distance	93.6
Height-curve error	87.2
Width-curve error	83.0
Area-curve error	83.0
Component-count error	88.6
Component exact match	88.6
Topology-violation rate	61.8
Object-missing rate	94.4
Geometry score	91.5
Trajectory score	93.6
Deformation score	85.1
Dynamics score	91.5
Integrity score	88.6
Soft-body total score	93.6

E Human Evaluation of Physical Metrics

E.1 Protocol

The development study compares Wan-2.2 TI2V-5B with CogVideoX-1.5-5B I2V on 200 inputs: 50 rigid-body, 50 soft-body, 50 liquid and 50 smoke cases. A single evaluator selects the video with more physically reasonable dynamics, judging plausibility rather than fidelity to the simulated reference. Agreement is the fraction of valid, non-tied pairs on which the metric and evaluator prefer the same video. Tables 9 , 10 and 11 report all supplied results.

E.2 Interpretation

Soft-body and liquid total scores achieve 93.6% and 98.0% agreement, respectively, while the reported rigid-body geometric metrics range from 71.4% to 81.6%. These results support agreement in this particular comparison, not general human alignment or superiority of aggregates over individual metrics. The study uses one evaluator and one model pair, and the human task assesses plausibility rather than reference fidelity.

F Evaluation Protocol Details

F.1 Material Score Definitions

The material scores combine complementary measurements of placement, deformation, and support evolution. Let $\phi(e;b) = 2^{-e/b}$ and $\psi(e;b) = e^{-e/b}$ for height-normalized errors; bars denote rollout averages. The rigid score is

$$S_{\text{rigid}} = 25\big[\bar{J} + c_m\phi(\bar{D}^{\text{fg}}_{\text{CD}};0.20) + c_m\phi(\bar{D}_{\text{ctr}};0.10) + c_p\phi(\bar{D}_{\text{ATE}};0.10)\big], \tag{8}$$

Table 10: Agreement of liquid metrics with human preferences. Tied and invalid pairs are excluded.

Metric	Agreement (%)
Mean IoU	98.0
Minimum IoU	93.9
Boundary Chamfer	96.0
Boundary HD95 (mean)	96.0
Boundary HD95 (maximum)	91.8
Boundary IoU ($0.005H$)	98.0
Boundary IoU ($0.01H$)	96.0
Occupancy FSS ($0.02H$)	98.0
Occupancy FSS ($0.10H$)	98.0
Temporal-change FSS ($0.02H$)	94.0
Temporal-change FSS ($0.10H$)	94.0
Centroid-trajectory error	94.0
Component-count error	90.0
Component exact match	91.8
Final-area error	89.8
Peak-area error	82.0
Largest-component-ratio error	92.0
Projected-area-curve error	90.0
Object-missing rate	96.8
Spatial score	98.0
Temporal score	94.0
Liquid total score	98.0

Table 11: Agreement of gas metrics with human preferences. Tied and invalid pairs are excluded. Threshold-dependent metrics use the `t025` configuration.

Metric	Agreement (%)
Mean IoU	93.6
Minimum IoU	95.3
Boundary Chamfer	93.6
Boundary HD95 (mean)	93.6
Boundary HD95 (maximum)	91.1
Boundary IoU ($0.005H$)	91.5
Boundary IoU ($0.01H$)	93.6
Occupancy FSS ($0.02H$)	91.5
Occupancy FSS ($0.10H$)	93.6
Temporal-change FSS ($0.02H$)	89.4
Temporal-change FSS ($0.10H$)	91.5
Centroid-trajectory error	85.1
Component-count error	42.6
Component exact match	60.5
Final-area error	71.1
Peak-area error	74.5
Largest-component-ratio error	83.0
Projected-area-curve error	80.9
Object-missing rate	100.0
Spread-RMS-curve error	89.5
Top-height-curve error	66.7
Spatial score	91.5
Temporal score	91.5
Smoke total score	91.5

where c_m is joint mask visibility divided by union visibility over actor–frame pairs, and c_p is jointly visible point–frame pairs divided by reference-visible pairs. Zero coverage makes the corresponding term zero; other unavailable required terms invalidate the score.

For the soft-body score in Equation 1,

$$\begin{aligned} \mathcal{G} &= \tfrac{1}{2}(\bar{J} + \bar{F}_\partial), \qquad \mathcal{D} = \sqrt{\psi(\bar{D}_{\mathrm{ctr}}; 0.10)\psi(E; 0.25)}, \\ \mathcal{I} &= \sqrt{[1 - \mathrm{clip}_{[0,1]}(r_{\mathrm{miss}} + r_{\mathrm{fp}})](1 - r_{\mathrm{comp}})}, \end{aligned} \tag{9}$$

where E averages the three relative-curve errors, r_{miss} and r_{fp} are missing and spurious-presence rates over all rollout frames, and r_{comp} is component-count mismatch conditional on both masks being visible. Complete presence mismatch gives $\mathcal{I} = 0$ even without component support; other non-finite required terms invalidate the score.

For the liquid score in Equation 2,

$$\begin{aligned} \mathcal{S} &= \tfrac{1}{3}\left[\bar{J} + \tfrac{1}{2}\sum_{\tau \in \{0.005, 0.01\}} \bar{B}_\tau + \tfrac{1}{2}\sum_{s \in \{0.02, 0.10\}} \overline{\mathrm{FSS}}_s\right], \\ \mathcal{T} &= \tfrac{1}{2}\sum_{s \in \{0.02, 0.10\}} \overline{\Delta\mathrm{FSS}}_s. \end{aligned} \tag{10}$$

The score requires visible, changing reference support and full reference-duration coverage. Soft and liquid sample scores average object totals only when all objects are valid.

For smoke at $\theta = 0.25$,

$$S_{\mathrm{smoke}} = 100\sqrt{\tfrac{1}{2}\left(\bar{J} + \overline{\mathrm{FSS}}_{0.10}\right)\overline{\Delta\mathrm{FSS}}_{0.10}}. \tag{11}$$

Full reference coverage, visible reference support, an observable reference change, and finite component means are required.

F.2 Essential Evaluation Settings

Alignment and averaging. The default mapping uses uniform scaling with relative aspect-ratio tolerance 10^{-3}. Masks and visibility use nearest-frame sampling on the reference clock; point coordinates use linear interpolation. Mask summaries exclude frame 0 and stop at the shared endpoint. Rigid mask metrics average finite actors within each frame, then frames; soft and fluid metrics average finite frames per actor, then actors. Point errors pool jointly visible point–frame pairs.

Geometry and spatial scales. Contours use 3×3 square erosion with zero exterior padding. Boundary Chamfer averages the two directed mean distances; HD95 takes the larger directed 95th percentile. Rigid foreground Chamfer sums the directed means over all foreground pixels. All distances divide by H. Boundary F1 uses an elliptical matching radius $\max(1, \mathrm{round}(0.01H))$. Boundary IoU uses inner bands from $\max(1, \mathrm{round}(\tau H))$ erosion iterations, $\tau \in \{0.005, 0.01\}$. FSS windows are sH, $s \in \{0.02, 0.10\}$, rounded to positive integers and increased by one if even, with zero exterior occupancy and fixed window area.

Point motion. Up to 20 queries per actor are sampled from the initial mask with seed 42 and three-pixel erosion, falling back to the original mask when necessary. Reference mesh points retain fixed barycentric coordinates and are projected through the rendering camera; available depth and object poses provide an alternative reference construction. CoTracker3 (Karaev et al., 2025) tracks the same queries in the prediction. D_{ATE} is mean Euclidean point error divided by H over jointly visible pairs; reference visibility checks projection, actor membership, and occlusion.

Presence and temporal support. Fluid presence requires $\max(4, \mathrm{round}(10^{-4}HW))$ foreground pixels; missing and false-positive rates divide by reference-present and reference-absent frame counts. Components use eight-connectivity, discarding areas below $\max(4, \mathrm{round}(10^{-3}A))$, with A the initial reference area for soft bodies and maximum reference area for fluids. For temporal FSS, the prediction index is $j(t) = \mathrm{round}(tf_P/f_G)$. Evaluate transitions only when this index changes, using the previous such observation (initially frame 0) as $a(t)$ in Equation 6.

Table 12: Liquid score under controlled reference-mask perturbations.

Perturbation	Draining tank	U-connector
Identity	100.0	100.0
Horizontal shift by $0.05H$	45.9	39.9
Delay by 1 frame	81.9	86.1
Delay by 4 frames	73.3	73.9
Frozen at frame 0	0.0	0.0

Empty observations. One empty mask gives zero overlap/F1/FSS and soft/fluid distance penalty $\sqrt{H^2+W^2}/H$; rigid distances require both masks. Both-empty masks are excluded from spatial comparisons, except rigid IoU equals one. Temporal FSS is excluded when neither support changes and is zero when exactly one changes. Undefined entries are excluded from finite means, with coverage reported.

Smoke and appearance. Smoke spread is $r(M) = \sqrt{|M|^{-1}\sum_{x\in M}\|x-\mathbf{c}(M)\|_2^2}/H$; onset is the first crossing of 1.1 times each side's rollout-start spread. We report spread-curve MAE and absolute onset-time error; absent or zero starting spread makes onset undefined, and unmatched crossings are missing or extra events. Declared occluders above opacity 0.5 are excluded on both sides. Temporal changes are compared only where both interval endpoints are observable. SSIM uses an 11×11 Gaussian window with $\sigma = 1.5$ and standard $C_1 = 0.01^2$, $C_2 = 0.03^2$ on RGB in $[0,1]$. The RGB temporal baseline is the negative mean squared difference between reference and predicted consecutive-frame RGB increments. RGB uses linear temporal interpolation for rigid scenes and nearest-frame sampling otherwise; rigid appearance includes frame 0.

F.3 Controlled Validation and Limitations

On two liquid references (1080p, 24 fps, frames 1–47), controlled perturbations show sensitivity to displacement, delay, and freezing (Table 12). Freezing retains spatial agreement near 0.7 but yields zero temporal agreement and total score, demonstrating the complementary role of the temporal factor.

Limitations. These image-space measurements depend on segmentation, tracking visibility, and temporal sampling. Earlier proxy-based development comparisons did not establish ranking superiority over simple baselines; the controls here demonstrate responses to prescribed perturbations.

G Domain-Specific Language

The DSL is organized into three layers, each controlling a distinct level of scenario randomization (Figure 5).

G.1 Object DSL

The Object DSL provides the benchmark's reusable object vocabulary, organized into three lists: *rigid objects*, *soft objects*, and *containers*. Each object is specified by its shape, physical material, and visual material, and can be instantiated by any entity regardless of its environment, camera, or layout (Figure 8).

Shape. Rigid and soft objects share the same seven primitives: ellipsoid, box, cylinder, cone, triangular prism, triangular pyramid, and square pyramid. Containers use a separate set of four shapes: box, cylinder, triangular prism, and hemisphere. Rope and cloth are soft matter but use segment count and grid resolution instead of primitive shapes, and are therefore defined directly by the Entity DSL.

Physical material. Rigid objects use densities of 500 or $2000\ \mathrm{kg/m^3}$. Soft objects are specified by named profiles that fix density, elastic modulus, Poisson ratio, and damping, avoiding per-scene

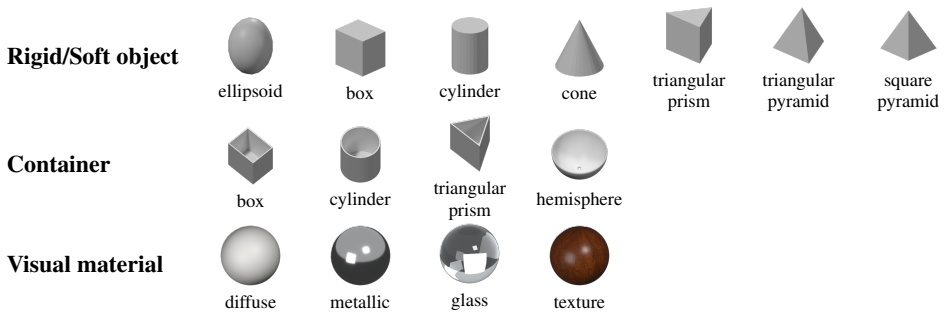


Figure 8: **Object DSL vocabulary.** Rigid and soft objects share seven shapes, while containers add four vessel shapes. Four visual materials are shown on a shared sphere, illustrating the independent specification of shape and appearance.

material tuning. Containers are static objects with fixed wall thickness, whose physical behaviour is determined by the solver.

Visual material. All three object lists support visual randomization through *diffuse*, *metallic*, and *glass* materials. Diffuse and metallic share a principled shader with sampled base colour, roughness, and metallicness; rigid objects use these with probability 0.6 and glass otherwise, while glass additionally specifies transmission and index of refraction. Containers may use non-metallic appearances to keep their contents visible, whereas soft objects remain opaque and exclude glass. The *texture* material is reserved for scene platforms, which use one of five measured assets—laminate floor, marble, plywood, rosewood veneer, or sandy gravel—instead of a sampled colour.

Entity-Level Liquid and Gas Liquids and gases lie outside the Object DSL: neither has primitive geometry or an independently sampled physical state. Their extent is determined by the containing vessel, displaced body, or emission source, while their material follows a fixed solver profile. They are therefore specified at the Entity DSL level through family-specific filling and emission parameters.

G.2 Entity DSL

Each of the twelve scenarios is documented by a card with three parts: a brief description of the scene and its target physical behaviour, a schematic of the released scene, and a table of entity-specific parameters listing every randomized quantity and its sampling interval or choices. The scenario cards are grouped by material class in Figures 10, 11, 12, and 13.

All schematics use a shared visual convention. Dashed rectangles denote objects in the Object List, while grey shapes denote entity-specific fixtures, such as platforms, ramps, poles, stands, tubes, and frames. The simulated matter itself (cloth, rope, liquid, or gas) is not an Object and is shown with a light fill and a label. Double-headed arrows indicate sampled or derived lengths and angles; thin dashed lines are construction guides; dash-dot lines denote family-fixed axes; solid arrows indicate driving motion; and curved arrows denote randomly sampled orientations.

Only entity-level randomization is annotated in the cards. Fixed quantities, such as stations, outlet bores, or fall gaps, are left out of the schematic as well as the table, since marking them adds symbols without adding randomization. Object shape, size, and visual material are randomized by the Object DSL, while camera viewpoints are specified by Scene DSL presets and are therefore excluded from the entity-specific parameters.

G.3 Scene DSL

The Scene DSL specifies the scene-level environment and camera. These are fixed throughout each 48-frame rollout, while being randomized or selected across scenes to provide controlled variation in appearance and viewpoint.

Environment. Each scene samples one of seven 4K equirectangular indoor HDRIs and uses it for both illumination and background: a ballroom, a music hall, a seaview suite, or one of four small-room variants (Figure 9). Using the same map for lighting and background keeps the appearance photorealistic and self-consistent, while fixing it throughout the rollout makes illumination a controlled variable.

Camera. Each rollout uses a static 50 mm perspective camera with a 32 mm sensor. Its pose is derived from the entity layout to keep the objects and their motion envelope in frame by construction. Container scenes fit their content within 72% of the horizontal and 75% of the vertical field of view, while line scenes use fixed offsets from the release line.

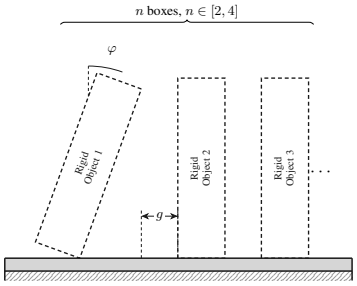
Viewpoint variation is restricted to discrete family-specific choices where different sides provide distinct observations: *dominoes* and *rolling_objects* use the front or back side of the family axis, while *gravity_draping_flag* uses the left or right side of its pole. Framing is validated before compilation, and layouts that exceed the visible motion envelope are rejected and redrawn.



Figure 9: **Environment maps.** The seven environment maps, each rendered as a standard HDRI probe: a chrome ball standing in the map used as the world background. One map is sampled per scene and serves as illumination and background for all 48 frames.

`dominoes` releases a line of 2–4 heavy boxes on one axis, all at rest. The first box is pre-tilted past its critical tipping angle and topples the chain; the remaining boxes stand upright until struck.

ENTITY VISUALIZATION

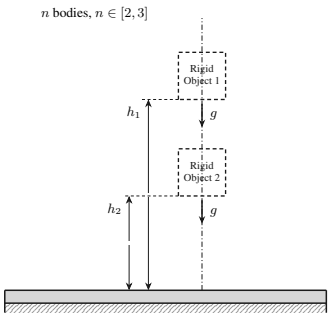


Entity-specific parameters

Symbol	Parameter	Sampling
n	number of boxes in the chain	integer in $[2, 4]$
φ	initial tilt of the first box	$20°$–$35°$, five degrees past its critical tipping angle
g	edge gap between consecutive boxes	$[0.12, 0.18]$ m

`falling_objects` releases 2–3 heavy rigid bodies from rest on one fixed vertical line. The bodies start at different heights and fall onto the platform, so the phenomenon is a free fall closed by a first contact and a subsequent impact chain.

ENTITY VISUALIZATION

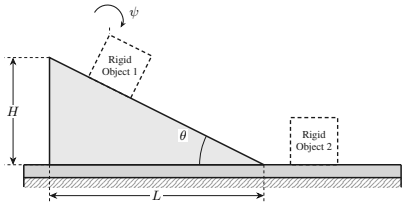


Entity-specific parameters

Symbol	Parameter	Sampling
n	number of falling bodies	integer in $[2, 3]$
h_i	release height of body i	its lower face is sampled uniformly in $[\ell_i,\ 1.5 - \ell_i]$ m, where ℓ_i is the half-height of body i

`rolling_objects` places exactly two heavy rigid bodies in one scene: one rests on an incline of sampled slope, one rests on the platform in front of it. Gravity drives the upper body down the slope into the lower one.

ENTITY VISUALIZATION



Entity-specific parameters

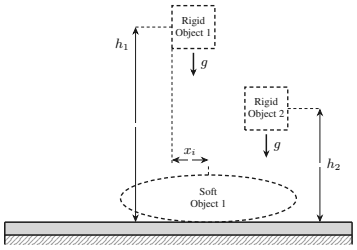
Symbol	Parameter	Sampling
L	ramp length	$[1.2, 1.8]$ m
H	ramp height	$[0.4, 0.8]$ m, with $H < L$ so the slope stays below $45°$
θ	ramp inclination angle	$\arctan(H/L)$
ψ	yaw of each body about its support axis	uniform in $(-180°, 180°]$, drawn independently for the two bodies

Figure 10: The three rigid scenarios. Every card gives the scene in words, the released scene with its entity-specific parameters, and the interval each randomised parameter is drawn from.

`falling_objects_on_soft_pad` releases one to three heavy rigid bodies from rest above a soft pad. The pad rests on the platform, the bodies fall onto it, and it deforms and returns, so the phenomenon is a soft–rigid impact closed by the elastic recovery of the pad.

ENTITY VISUALIZATION

n bodies, $n \in [1, 3]$



Entity-specific parameters

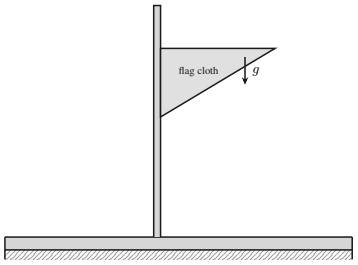
Symbol	Parameter	Sampling
n	number of falling bodies	integer in $[1, 3]$
h_i	release height of body i, its centre above the platform	uniform in $[1.20, 1.60]$ m
x_i, y_i	horizontal position of body i over the pad	uniform inside the pad footprint, keeping 0.06 m to its edge and 0.02 m between bodies

`gravity_draping_flag` holds a cloth in a vertical plane along one edge of a fixed pole and releases it from rest. Gravity pulls the free edge down, so the phenomenon is the draping and the wrinkling of the cloth as it settles against the pole.

ENTITY VISUALIZATION

cloth outline: one of four (see the table)

pole side: left ↔ right, the cloth hangs on that side



Entity-specific parameters

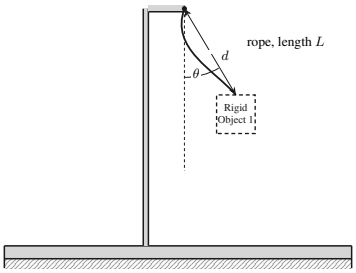
Symbol	Parameter	Sampling
cloth outline	outline of the cloth	right triangle, rectangle, centered trapezoid or right trapezoid, each with weight $1/4$
cloth size	defining lengths of that outline	height in $[0.70, 0.95]$ m with a width-to-height ratio in $[1.20, 1.70]$
cloth colour	hue, saturation and value of the cloth	hue in $[0°, 360°]$, saturation in $[0.45, 1.00]$, value in $[0.30, 0.85]$
pole side	side of the pole the cloth hangs from	left or right

`gravity_tightening_soft_rope` ties a soft rope to a fixed anchor and releases a heavy load from rest closer to the anchor than the rope is long. The rope starts slack and is pulled straight as the load falls, so the phenomenon is the tightening and the swing that follows.

ENTITY VISUALIZATION

release: the load is dropped from rest

$d < L$, so the rope hangs slack at release



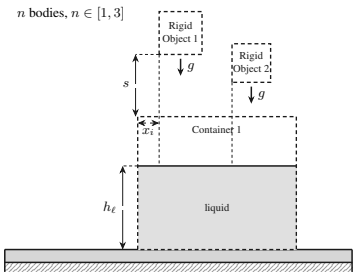
Entity-specific parameters

Symbol	Parameter	Sampling
L	rest length of the rope	uniform in $[0.8, 1.4]$ m
ρ	fraction of the rope length at which the load is released	uniform in $[0.55, 0.90]$
d	distance between the anchor and the load at release	ρL, so the rope starts slack
θ	angle of the load from the vertical at release	uniform in $[-45°, 45°]$
rope colour	colour and finish of the rope	each colour channel in $[0.08, 0.95]$, roughness in $[0.20, 0.80]$, metallicness in $[0, 0.35]$

Figure 11: The three soft scenarios, on the same three-part contract as Figure 10.

`rigid_objects_in_water_tank` fills a glass container with still liquid and releases one to three heavy rigid bodies from rest above it, so the phenomenon is buoyancy and the fluid–solid displacement of a free surface.

ENTITY VISUALIZATION

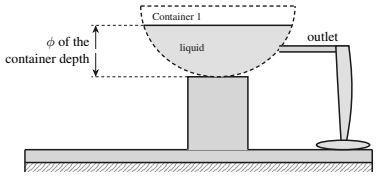


Entity-specific parameters

Symbol	Parameter	Sampling
n	number of falling bodies	integer in $[1, 3]$
h_ℓ	still depth of the liquid, from the container floor	uniform in $[0.35, 0.65]$ m
s	drop height of each body above the container mouth	uniform in $[0.12, 0.34]$ m
x_i, y_i	horizontal position of fall path i in the cross-section	uniform over the container cross-section, keeping 0.03 m to the wall and 0.02 m between paths
liquid colour	hue, saturation and value of the liquid	hue in $[0^\circ, 360^\circ]$, saturation in $[0.10, 0.35]$, value in $[0.78, 0.95]$

`draining_tank` fills a glass tank that stands on a support and is open at one side, so the phenomenon is an emptying tank: a falling free surface, an outflow and the pressure that drives it.

ENTITY VISUALIZATION

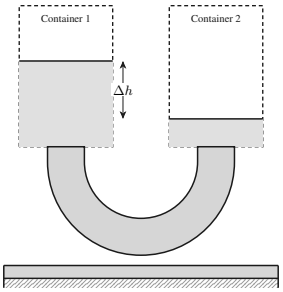


Entity-specific parameters

Symbol	Parameter	Sampling
ϕ	fraction of the container depth covered by the liquid	uniform in $[0.72, 0.80]$ from the floor, so the liquid never reaches the outlet at rest
liquid colour	hue, saturation and value of the liquid	hue in $[0^\circ, 360^\circ]$, saturation in $[0.10, 0.35]$, value in $[0.78, 0.95]$

`u_connector` joins two glass containers with a U tube. The liquid is initialized at unequal levels and then moves toward equal levels under gravity.

ENTITY VISUALIZATION



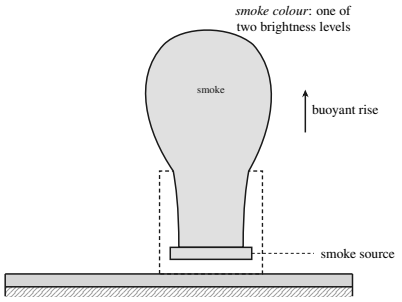
Entity-specific parameters

Symbol	Parameter	Sampling
Δh	signed initial liquid-level difference, defined as $h_L - h_R$	magnitude spans 30%–90% of the tighter container's usable freeboard; either sign is equally likely
V	initial volume of liquid in the scene	uniform in $[35\%, 95\%]$ of the geometry-derived valid interval

Figure 12: The three liquid scenarios, on the same three-part contract as Figure 10.

`smoke_stack` injects smoke into the lower fifth of a metallic container. The gas is lighter than the air around it, so the phenomenon is a buoyant plume: it rises through the cavity, leaves the open top and spreads above the rim.

ENTITY VISUALIZATION

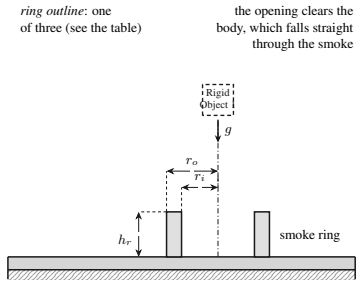


Entity-specific parameters

Symbol	Parameter	Sampling
smoke colour	hue, saturation and value of the plume	hue in $[0°, 360°]$, saturation in $[0.02, 0.10]$, value either 0.15 or 0.92

`impact_smoke_dispersion` rests a still smoke ring on the platform and releases a heavy rigid body above its opening on the same axis. The body does not touch the smoke: it drives the air ahead of itself, so the phenomenon is a vortex ring that disperses as the body falls through it.

ENTITY VISUALIZATION

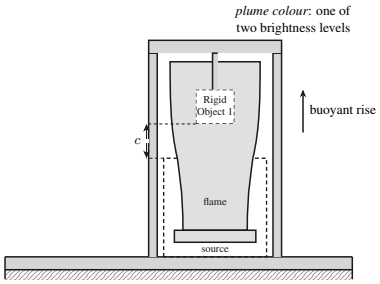


Entity-specific parameters

Symbol	Parameter	Sampling
ring outline	outline of the static ring	annulus, rectangular frame or triangular frame
r_o	outer extent of the ring	half-width in $[0.24, 0.32]$ m
r_i	inner extent of the opening	r_o minus a band of $[0.035, 0.06]$ m, so the body gets through
h_r	height of the ring	in $[0.12, 0.18]$ m
smoke colour	hue, saturation and value of the ring	hue in $[0°, 360°]$, saturation in $[0.02, 0.10]$, value 0.15 or 0.92

`flame_heating` burns a reacting gas inside a metallic container and holds a rigid body on a frame above the opening, so the phenomenon is a hot plume split by the body and closing over it.

ENTITY VISUALIZATION



Entity-specific parameters

Symbol	Parameter	Sampling
c	clearance from the container rim to the lower face of the body	in $[0.088, 0.35]$ m, so the plume is squeezed through the gap
plume colour	hue, saturation and value of the plume	hue in $[0°, 360°]$, saturation in $[0.02, 0.10]$, value 0.15 or 0.92

Figure 13: The three gas scenarios, on the same three-part contract as Figure 10.

H FINE-TUNING DETAILS

This section provides the implementation details of the Wan 2.2 TI2V 5B fine-tuning experiment described in the main paper, including the training data, LoRA configuration, optimization settings, and evaluation protocol.

H.1 TRAINING DATA

We construct the fine-tuning set from two soft-body simulation scenarios: `falling_ball_on_soft_cube`, which is used only for this fine-tuning study and is not included in the released dataset, and `hanging_soft_rope`. Each scenario contains 128 videos, resulting in 256 videos in total.

Notably, the evaluation on `falling_objects_on_soft_pad` is conducted on a different soft-body scenario from those used for fine-tuning. The substantial performance improvement observed on this scenario suggests that the adaptation is not limited to reproducing the exact training setup, but can transfer to related soft-body interactions with different scene configurations.

We independently split each scenario using a fixed random seed (`seed`=42) and a training fraction of 0.7. This results in 90 training videos and 38 held-out videos per scenario, or 180 training videos and 76 held-out videos in total. The training and held-out subsets contain no overlapping instances.

Each training video is associated with a text prompt and a scenario identifier. Since Wan 2.2 TI2V is an image-conditioned video generation model, the first frame of each video is used as the image condition.

We use one fixed prompt for each scenario:

> *A rigid sphere falls under gravity onto the soft cube, compressing and deforming the cube while the camera remains fixed.*

and

> *A soft rope fixed at its upper end swings and deforms under gravity while the camera remains fixed.*

During training, videos are processed at a resolution of 640×352 with 45 frames per sample. We set `dataset_repeat`=10, which repeats the 180 unique training videos within each epoch rather than introducing additional independent samples.

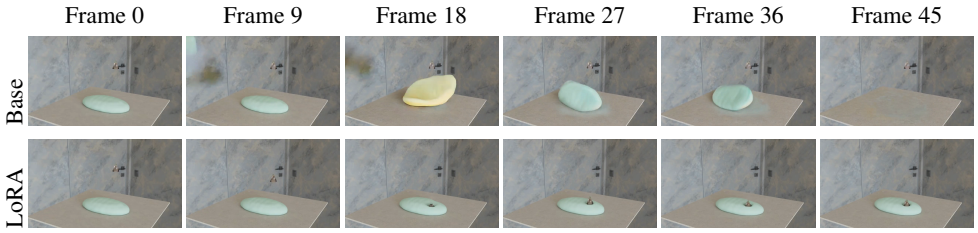


Figure 14: Qualitative comparison between the base Wan 2.2 TI2V 5B model and the LoRA fine-tuned model on a soft-body scene. Each column shows the same temporal frame for the two models.

H.2 LORA IMPLEMENTATION

We fine-tune Wan 2.2 TI2V 5B using Low-Rank Adaptation (LoRA), implemented with DiffSynth-Studio and Hugging Face Accelerate under the standard flow-matching objective. Rank-32 adapters target the diffusion transformer's attention projections (`q`, `k`, `v`, `o`) and feed-forward layers (`ffn.0`, `ffn.2`), while all pretrained parameters remain frozen.

Training starts from the pretrained checkpoint without adapter warm-starting or checkpoint resumption. Gradient checkpointing reduces memory consumption. We save only adapter parameters;

Table 13: Inference configurations of the evaluated video generation models.

Model	Resolution	Frames	FPS	Steps	CFG
Wan2.2-TI2V-5B	1024×576	49	24	30	5
Cosmos-Predict2.5-2B	1024×576	33	16	36	7
Cosmos3-Edge	768×432	49	24	20	6
HunyuanVideo-1.5	1280×720	49	24	12	1
CogVideoX1.5-5B-I2V	512×288	49	16	30	6
Wan2.1-I2V-14B	1280×720	41	16	40	5
Wan2.2-I2V-A14B	1280×720	41	16	40	3.5

inference therefore requires both the pretrained model and the trained adapters. Figure 14 compares the base and adapted models using time-aligned frames, illustrating differences in predicted motion and deformation on a soft-body scene.

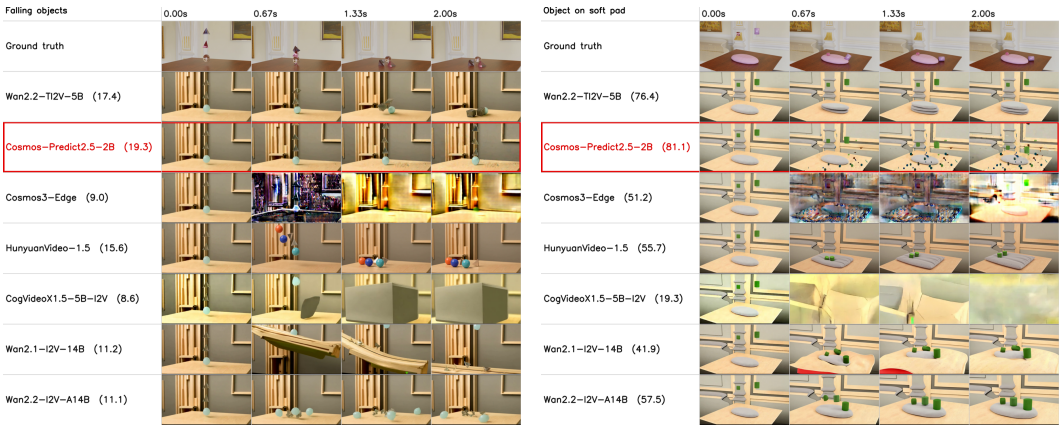
H.3 Training Configuration

We train on four NVIDIA A800-SXM4-80GB GPUs using distributed data parallelism with Hugging Face Accelerate. The per-device batch size is one, with one gradient-accumulation step, giving an effective global batch size of four. We use AdamW with a learning rate of 10^{-4} and weight decay of 0.01. The learning-rate scheduler is configured as `ConstantLR`. Training uses BF16 precision; model and optimizer CPU offloading, FP8, and quantization are disabled.

We train for ten epochs. With 180 distinct training videos and `dataset_repeat`=10, each epoch presents 1,800 samples and comprises 450 optimizer steps. The complete run therefore comprises 4,500 optimizer steps. LoRA checkpoints are saved every 100 steps, and the final `step-4500.safetensors` checkpoint is used for evaluation.

I Inference Configurations and Qualitative Comparisons

We detail the inference configurations of the seven evaluated models, including resolution, frame rate, sampling steps, and guidance scale. Each model uses its official implementation and native settings without benchmark-specific fine-tuning in Table 13. We also present representative frames for qualitative comparison on one scene per material category—rigid bodies, soft bodies, liquids, and gases—under identical initial-image and prompt conditions. These examples illustrate differences in predicted dynamics and complement the quantitative evaluation in Figure 15.



(a) Rigid: falling objects.

(b) Soft: object falling on a soft pad.

(c) Liquid: draining tank.

(d) Gas: smoke stack.

Figure 15: Qualitative comparison of the seven evaluated models on one scene per material category. Each block shows the ground-truth rollout in the top row and the seven generated rollouts below it; the four columns are frames at matched reference times (0.00–2.00 s). The best-scoring model of that scene is highlighted in red.