SCENEPROBE: RECONSTRUCTING AND PHYSICALLY TESTING SCENES FROM IMAGES AND VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

A scene reconstructed from an image or video must withstand gravity and controlled actions, yet a plausible render cannot show whether its objects move independently, a hinge is present, a cushion has a filled volume, or transient collisions occur. We present SceneProbe, an image/video-to-MuJoCo pipeline that keeps each source instance linked to a native physical component and tests the exact exported model. Its Instance Evidence Graph records masks, geometry, placement and support/part cues; Type-Aware Native Assembly creates separate rigid bodies, fitted joints and, when feasible, contact-enabled tetrahedral objects with declared parameter priors; and Intervention-Aware Native Audit checks realized structure and every integration state under gravity and forces. In a selected corrected living-room image scene, 13 predicted targets include two functional hinges and two volumetric pillows, and the five-second rollout reaches 1.825 mm peak enabled-contact penetration. Corrected toy tabletop image and video scenes also run, although the video route remains mainly first-frame conditioned. Across 68 dependent development XMLs, saved 10 Hz states miss 14 of 38 every-step contact violations at a 2 mm operational threshold. None of ten frozen public-image first attempts passes the combined coverage, completion and contact gate; four authored source-truth RGB scenes also fail complete structure/action recovery. SceneProbe demonstrates auditable native reconstructions in selected cases while exposing an automation gap. Mass, friction and elasticity are priors, not recovered real-world measurements.

1 INTRODUCTION

An image can depict a convincing room while leaving a simulator's essential choices unresolved: which objects move independently, what supports them, which parts have joints, and which surfaces bound a deformable volume. Mass, friction and elasticity are still less observable. Yet an executable scene needs these choices to train robots, test interventions or support interactive editing. We therefore treat image/video-to-simulation as constructing a physical *hypothesis* whose native components and response can be inspected, rather than as matching one camera view.

Image and video reconstruction systems now produce strong object geometry and editable scenes (Chen et al., 2026; Ranawaka et al., 2026; Xia et al., 2026a; Zhang et al., 2026b; Lee et al., 2026; Ma et al., 2026). The difficult remaining interface is between predicted objects and a runnable, typed scene: a plausible mesh may hide an absent hinge, an empty cloth shell may stand in for a filled cushion, and saved render frames can miss brief collisions. A gravity-stable scene can also respond incorrectly to an action.

We introduce **SceneProbe**, which carries the same object IDs through three modules. The *Instance Evidence Graph* records masks, lifted geometry and proposed support/part relations. *Type-Aware Native Assembly* exports separate bodies, supported joints and feasible tetrahedral volumes with explicit parameter priors. *Intervention-Aware Native Audit* reopens the exact MuJoCo XML and tests structure, every-step contact and controlled response. Figure 2 uses recorded source, intermediate and native results; Figure 1 shows selected reconstructed scenes.

The selected living-room image yields 13/13 *predicted* assets, two hinges and two volumetric pillows, with 1.825 mm peak contact penetration during five seconds of gravity. A selected toy-

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: Three distinct selected *image-input* scenes. Each row shows source RGB, the exact MuJoCo input-view scene and a later native frame. The living-room XML has separate rigid objects, two hinges and two tetrahedral pillows; the last panel follows a 2 N pillow-force probe on that XML. Toy and bathroom rows show five-second gravity rollouts. The bathroom is explicitly an assisted *rigid fallback*: four proposed soft assets became rigid. Counts refer to predicted inventories, and physical constants are priors rather than measured properties. A toy-video route to the same toy scene appears in the supplement.

video scene reaches 1.929 mm, but remains mainly first-frame conditioned. The limits are as informative as the demos: in 68 dependent development XMLs, 10 Hz state inspection misses 14 of 38 every-step contact violations; all ten frozen public-image first attempts fail the joint coverage/completion/contact gate, and four authored source-truth scenes fail complete structure/action recovery. Physical parameters remain priors, not identified measurements.

Our contributions are (1) an identity-preserving representation and native assembly of heterogeneous predicted assets; (2) an exact-export audit that separates coverage, type, contact and response; and (3) failure-inclusive tests that expose what selected demonstrations and support/contact corrections do and do not establish.

2 Related work

2.1 Object-centric reconstruction from images and video

SAM 3, SAM 3D and MoGe provide prompted masks, object geometry and monocular geometric cues (Carion et al., 2025; SAM 3D Team, 2025; Wang et al., 2024); CAST and MIDI study composition from one view (Yao et al., 2025; Huang et al., 2025). FIRE3D predicts separate, interaction-ready objects from an RGB image or casual video (Xia et al., 2026b). OVOW, SimRecon and ReplicateAnyScene exploit video for instance geometry or composition (Chen et al., 2026; Xia et al., 2026a; Dong et al., 2026), while Vid2Sim fits appearance, geometry and physical properties to observed video motion (Chen et al., 2025b). Our video route merely groups sampled masks and lifts geometry mainly from frame zero; it neither recovers full 4D dynamics nor iden-

tifies material constants. SceneProbe inherits perception components and follows predicted object IDs into one inspectable native model. Its audit addresses realized joints, colliders and response, which novel-view or mask scores alone cannot certify.

2.2 Simulation-ready scenes and physical asset models

Single-image interactive scene systems include PhysGen3D and WonderPlay, which simulate image-conditioned dynamics under user actions (Chen et al., 2025a; Li et al., 2025). PhysOmni, SimuScene, REST3D, Picasso and ϕ-Scene assemble or physically refine multi-object scenes (Zhang et al., 2026b; Lee et al., 2026; Ma et al., 2026; Yu et al., 2026; Li et al., 2026); SimFoundry and RoboSnap pursue editable video-derived or robot-oriented scenes (Ranawaka et al., 2026; Zhang et al., 2026a). TabletopGen and TableVerse address dense interactive tabletop layouts (Wang et al., 2026c;a). At the asset level, PhysX-Omni and ManiTwin produce simulation-ready objects (Cao et al., 2026; Wang et al., 2026b); Real2Code, Articulate-Anything, URDF-Anything+ and SPARK model executable parts and joints (Mandi et al., 2024; Le et al., 2025; Wu et al., 2026; He et al., 2025). REACT3D adds articulations to static 3D scenes (Huang et al., 2026). These works establish image/video-to-simulation and mixed physical assets as existing directions. SceneProbe records the identity-to-native mapping and checks the exact compiled structure, but its heuristic part assembly misses authored hinges and slides. We have no matched external-system advantage or downstream policy result.

2.3 Simulator-level validation and intervention

MuJoCo exposes native bodies, contacts, joints and deformable state for repeatable interventions (Todorov et al., 2012). PhysSplat estimates property distributions and simulates motion in reconstructed 3D scenes; Splatting Physical Scenes refines geometry and physics from robot sequences (Zhao et al., 2025; Moran et al., 2025). VR-Robo and GS-Playground connect reconstructed or photorealistic scenes to robot simulation (Zhu et al., 2025; Jia et al., 2026). Relative to these systems, our parameters are unmeasured priors and our contribution is a paired diagnostic on the exact XML: initial, 10 Hz and every-step contact checks plus authored action-response tests. The 68 XMLs are dependent development variants, and the 2 mm threshold is operational rather than physical ground truth. The audit complements visual, geometric and robot-task evaluation; it does not establish simulator-to-reality agreement.

3 Method

Problem definition. Let $X = \{I_t\}_{t=0}^{T-1}$ be a monocular RGB input, with $T = 1$ for an image. We seek an executable scene $S = (\mathcal{B}, \mathcal{G}, \mathcal{J}, \mathcal{F}, \Theta)$, where $\mathcal{B}$ is a set of independently addressable bodies, $\mathcal{G}$ their visual and collision geometry, $\mathcal{J}$ their internal joints, $\mathcal{F}$ any deformable finite elements, and Θ the simulator parameters. The desired set of objects is the visible foreground plus selected load-bearing supports; background pixels are outside the declared inventory. A single uncalibrated view does not identify mass, friction or elasticity. Accordingly, Θ contains explicit category/geometry priors unless independent physical measurements exist; the output is a testable physical hypothesis rather than recovered ground truth.

Pipeline overview. SceneProbe maintains the same object ID from source evidence to a native simulator component and then to an audit finding. The *Instance Evidence Graph* binds masks and first-frame geometry to common-frame placement, support parents and part cues. *Type-Aware Native Assembly* maps each graph node to its own fixed/free body, fitted joint assembly or filled soft volume, and exports the exact XML with referenced meshes. *Intervention-Aware Native Audit* compiles that XML, checks the realized structure, and tests gravity and controlled forces while inspecting every integrated state. A failed audit may guide a separately saved *selected development derivative*; its outcome does not replace the initial export. Figure 2 makes these interfaces and the selection boundary visible. The modules are implemented decision rules around inherited perception backbones, rather than newly trained estimators.



Figure 2: SceneProbe on one selected image. Left: saved RGB, predicted masks and depth. Center: four saved instance crops become ID-linked meshes with scale, anchor pose and a table-support parent; the lower image is the actual assembled MuJoCo render. Native rigid, hinged and filled-tetrahedral representations receive explicit parameter priors. Right: native response after a 2 N pillow load, with hinge and pillow traces from *separate* probes of the same XML. This selected development derivative does not establish first-attempt success or real material identification.

3.1 Instance Evidence Graph

We represent the predicted scene as $\widehat{H} = (V, E_s, E_k)$. A node

$$v_i = (\mathrm{id}_i, c_i, \{M_{i,t}\}, \widehat{g}_i, \widehat{p}_i, \tau_i, \theta_i^0) \tag{1}$$

contains a persistent ID, category c_i, sampled image masks $M_{i,t}$, lifted geometry $\widehat{g}_i$, common-frame placement $\widehat{p}_i$, proposed physical type τ_i, and an uncalibrated parameter prior θ_i^0. Types distinguish fixed, free rigid, articulated, filled soft and surface-cloth targets. A directed edge $(i, j) \in E_s$ states that child i is predicted to rest on support j; E_k holds optional tuples $(i, p, \widehat{a})$, where p is a candidate within-object moving part and $\widehat{a}$ is its proposed joint axis. Parts need not be separate scene nodes. A vision-language review proposes the visible inventory, types and edges; SAM 3 estimates prompted masks (Carion et al., 2025); SAM 3D lifts per-object textured geometry (SAM 3D Team, 2025); and MoGe depth provides common-frame cues (Wang et al., 2024). These are inherited estimators, and their uncertain predictions remain attached to the graph.

For a support-mask pixel $u = (x, y)$ with homogeneous coordinate $\widetilde{u} = (x, y, 1)^\top$, predicted metric depth $z(u)$ and estimated camera intrinsics K, the point used for support fitting is

$$p^w(u) = W\,[z(u)K^{-1}\widetilde{u}; 1], \tag{2}$$

where W is the camera-to-scene transform inferred from the predicted support plane. This aligns objects to a shared scene frame; it does not measure hidden geometry or independent real-world scale. For video, sampled SAM 3 tracks are grouped into reviewed whole-object IDs using first-frame mask purity, box overlap, category agreement, and a spatial check for later-only tracks. Geometry and initial pose are then generated principally from frame zero. The present temporal evidence groups masks but does not recover a time-varying physical state or material constants; stationary appearance alone does not make an object world-fixed. The saved masks, source-frame associations and missing-node decisions let an auditor trace a failed native object back to the corresponding observation.

3.2 Type-Aware Native Assembly

The assembly map uses the same graph to create $S = A(\widehat{H}; \Theta_{\mathrm{prior}})$. For each node v_i, it assigns a separate native body or body assembly $B_i \subset \mathcal{B}$, meshes and collision geometry $G_i \subset \mathcal{G}$, and prior

parameters $\theta_i \subset \Theta_{\text{prior}}$. Fixed supports become world-attached bodies with finite support colliders; free rigid targets receive movable bodies, convex collision proxies, mass and friction priors. An articulated target receives parent and child bodies with a hinge only if a moving part and rotation axis can be fitted; an unsupported slide is recorded as absent. A threaded closure is not reinterpreted as a valid pure hinge. For a predicted filled soft target, a repaired closed surface is tetrahedralized to produce elements $F_i \subset \mathcal{F}$ with contact enabled. A surface cloth shell and a filled tetrahedral solid are recorded as different realized types; if meshing fails, any rigid fallback is reported as a type change. The assembled model is therefore inspectable through the identity-preserving map

$$v_i \longmapsto (B_i, G_i, J_i, F_i, \theta_i), \qquad J_i \subseteq \mathcal{J},\ F_i \subseteq \mathcal{F}. \tag{3}$$

Here J_i contains native internal joints, if any. An upstream instance enters the map only when both its scaled mesh and anchor pose exist; otherwise it is recorded as missing. A free rigid body's mass follows $m_i = \rho_i \operatorname{vol}(G_i)$, where ρ_i is the category-conditioned density prior and $\operatorname{vol}(G_i)$ is estimated from its generated shape. Friction, restitution and soft time constants are likewise stored material-class priors, not visually identified constants. The code never estimates an elastic modulus by fitting observed deformation.

A support mask and the points in Eq. (2) determine a finite collision surface for its predicted children. The later *top-cluster* development variant restricts valid support-point heights to their upper 75%, finds the densest 5 mm height bin, and retains points within 8 mm of that mode. It accepts a fitted plane only if at least 20% of mask points remain, its 90th-percentile residual is at most $\max(2\,\text{mm}, 0.02D)$, and its tilt is below 25°, where D is the selected point-cloud extent. Its projected convex boundary becomes a bounded support slab. These thresholds were chosen after seeing source-case failures and are not the cold-export default. Where raw RGB would imprint foreground objects into a support texture, an object-removed LaMa image supplies the projected texture (Suvorov et al., 2022); this changes appearance, not physical evidence.

An optional, separately reported development repair moves whole free bodies using the *compiled native collision geometry*. Let $\Delta_i \in \mathbb{R}^3$ be the translation of root i, ℓ_i its collision-bounding diagonal, and Δ_i^{seat} an optional shift onto its predicted support. The refiner approximately solves

$$\min_{\Delta}\ \tfrac{1}{2}(\Delta - \Delta^{\text{seat}})^\top Q(\Delta - \Delta^{\text{seat}}) \quad \text{s.t.} \quad a_k^\top \Delta \geq b_k, \quad |\Delta_{i,d}| \leq \alpha \ell_i, \tag{4}$$

Here $\Delta \in \mathbb{R}^{3n}$ stacks root shifts, $Q \succ 0$ is a diagonal size-weighted displacement penalty, $(a_k, b_k) \in \mathbb{R}^{3n} \times \mathbb{R}$ linearizes a native contact separation, d indexes translation coordinates, and α is the recorded per-coordinate shift cap (0.25 by default; 0.30 in the selected top-cluster derivative). Contacts are recomputed and constraints retained across iterations. Internal joints, geometry, gravity and material parameters remain fixed. The derived XML is recompiled and its world geometry checked against the optimized state; a low initial overlap does not imply a successful rollout. No source-truth masks or poses enter this refiner. Appendix B records the variants and their selection status.

3.3 Intervention-Aware Native Audit

The auditor loads the exact exported XML in MuJoCo (Todorov et al., 2012) and inspects the realized components in Eq. (3). Let $n = |V|$ and let $r_i = 1$ when target i has a source geometry file, visual geometry and native representation, and $r_i = 0$ otherwise. We report internal predicted-inventory coverage $C_{\text{pred}} = n^{-1} \sum_i r_i$; this cannot measure recall of objects omitted from V. Type-consistent counts additionally require J_i or F_i when specified. For an intervention sequence $U = \{u_t\}_{t=0}^{L-1}$, the native transition produces $q_{t+1} = f_S(q_t, u_t)$, where q_t is the full simulator state (positions, velocities and deformable state), f_S is its discrete integrator, and u_t is the prescribed control or disturbance, zero for gravity-only tests. We inspect the initial state and every state after integration/forward evaluation. If $d_{c,t}$ is MuJoCo's signed distance for enabled contact c at state t, and $V_{e,t}$ is signed volume of deformable tetrahedron e, our diagnostics are

$$P_{\max} = \max_{t,c}\big[\max(0, -d_{c,t})\big], \qquad J_{\min} = \min_{t,e} \frac{V_{e,t}}{V_{e,0}}. \tag{5}$$

The extrema range over audited discrete times t and active contacts/elements; an empty contact set has $P_{\max} = 0$. $V_{e,0} \neq 0$ is the initial signed volume, so $J_{\min} \leq 0$ flags collapse or orientation

reversal; $J_{\min}$ is not applicable without tetrahedra. Surface cloth instead uses triangle-area collapse checks. We also require the requested rollout duration, finite states and no simulator warning. Our development contact target is $P_{\max} < 2\,\text{mm}$, not a calibrated physical-validity theorem. A sparse comparator computes the same maximum over the saved 10 Hz subset of states, isolating the inspection cadence while holding S and U fixed. Where an authored source supplies a prescribed action and reference trajectory, we also compare the candidate's driven part/target response after settling; no such ground-truth response exists for the natural RGBs. Ground acceleration, joint and soft forces are native probes; camera shake in demo videos is a post-render effect and excluded from physical metrics. Collision exclusions and discrete sampling bound what this audit can certify.

4 EXPERIMENTS

4.1 DATA AND IMPLEMENTATION

We evaluate distinct strata rather than pooling unlike inputs. Selected development demonstrations include a living-room image, toy tabletop image and video, and other local scenes with separately recorded correction histories. Eight additional local images and three ordered public images retain all original first attempts. A separate ten-image sequence uses numbered RGBs 04–12 and 15 from a public scene-project page; its 60-entry visible-asset list is author written and is not exhaustive ground truth. Four related, authored MuJoCo scenes provide initial RGBs, source-visible masks, exact target types and prescribed actions for controlled truth-render-reconstruct-act tests. Finally, a retrospective audit cohort contains 68 distinct XML hashes from 27 development directories; several are derivatives of the same input. The public sequences are convenience samples, not matched official baseline evaluations. Supplementary Material gives input provenance, failure stages and frozen denominators.

All runs use MuJoCo 3.3.7. A reviewed Qwen vision-language inventory, SAM 3 masks, a SAM 3D adapter and MoGe geometry cues feed the native exporter; LaMa fills selected support textures. These are pretrained components: SceneProbe adds no model training, training stages or batch size. The image driver uses seed 42; the toy clip is sampled at 3 frames/s and retains mainly first-frame geometry. Rigid scenes use 2 ms integration, while the volumetric living-room scene uses 0.5 ms. The primary joint gate requires predicted-target coverage, a warning-free finite five-second gravity rollout, no invalid tetrahedra and peak enabled-contact penetration under 2 mm across every discrete state. We additionally report predicted-mask input-view IoU as an internal alignment proxy and prescribed-action error only for authored sources. Same-input contact, support, floor and inspection-cadence ablations use exact XML lineages; post-outcome selected derivatives never replace original first-attempt counts. Full protocols, thresholds, commands, hashes and material priors are in the supplement.

4.2 NATIVE AUDIT AND RESPONSE DIAGNOSTICS

What sparse checks miss. In the retrospective 68-XML corpus, all 68 five-second simulations completed without MuJoCo warnings *by selection*, but the every-step contact audit finds 38 XMLs with peak penetration at least 2 mm. The 10 Hz saved-state check finds only 24 of these 38 and misses 14 peak-contact violations between recorded frames. This is a paired comparison on saved development trajectories, including multiple derivatives per input; it does not estimate a population miss rate or show a reconstruction improvement. Separately, a stronger synthetic acceleration run has no MuJoCo warning and no inversion in its 20 saved render states, yet the every-step audit finds one inverted pillow tetrahedron at 1.6045 s and stops the run. A duration/completion check also flags that run; the element audit identifies the physical failure.

At 2 mm, 14 of the 38 violations are present at initialization and 24 first appear later; 10 of those later violations appear at 10 Hz and 14 do not. Controlled simulator-truth probes show a complementary gap: 32 mass/friction variants pass contact but fail a fixed-force response tolerance, while 16 embedded variants fail contact but pass the post-settling response gate. A separate source-only damping control has 12/12 gravity-contact passes and 8/12 joint-response failures. These are dependent synthetic controls, not real-world task results (Appendices F and H).

Contact-depth threshold (mm)	1	2	3	5	10
Every-step violations / 68	59	38	27	17	10
Detected at initialization / 68	16	14	13	9	7
Detected at 10 Hz / 68	35	24	20	10	7
Missed at 10 Hz / 68	24	14	7	7	3

Table 1: Threshold sensitivity on the same retrospective, non-independent development XMLs. Each column changes only the diagnostic contact-depth threshold; it is not a new simulation experiment.

Same-input development lineage	Before (mm)	After (mm)	Remaining limitation
Toy image: support/contact	35.830	1.690	Selected derivative
Toy video: support/pair contact	55.867	1.929	Mostly frame-zero geometry
Bathroom: rigid-only to tuned rigid	68.925	1.621	Four soft types replaced
Authored chest: fitted support	117.805	1.543	Hinge absent
Authored drawer: fitted support	71.728	0.991	Slide absent
Authored soft pad: fitted support	82.372	0.612	Volume absent

Table 2: Five-second maximum contact penetration before and after a selected *post-outcome* correction on the same input. All values are from native MuJoCo audits; rows are related development cases and are not a frozen generalization estimate. The bathroom baseline is already a rigid fallback, and the authored rows still fail complete type/action recovery.

4.3 Reconstruction and structure recovery

Source-truth reconstruction tests. Four related authored MuJoCo scenes supply an initial RGB, target masks and an action reference for a laptop hinge, chest hinge, drawer slide or soft pad. The original candidates uniquely recover 4/4, 2/4, 2/3 and 2/4 source targets, respectively, but miss every required hinge, slide or volumetric soft structure. Their category-matched source-mask macro IoUs are 0.457, 0.338, 0.435 and 0.276. All four therefore fail complete structure/action recovery. Source impacts themselves reach 3.252–6.011 mm contact depth, showing that the 2 mm development target cannot be treated as calibrated task success. The appendix details the target associations, corrected authored hinge units and action probes (Appendix G).

Fixed support correction after outcome inspection. A shared image/depth top-cluster plane and seating cap was developed after seeing the first authored failure. It lowers the chest, drawer and pad five-second contact peaks from 117.805, 71.728 and 82.372 mm to 1.543, 0.991 and 0.612 mm, yet all still lack their required joint or volume. Applied after outcome inspection to ten public RGBs, it produces 0/10 contact passes; five derivatives complete and all exceed 2 mm. A corrected laptop source-range and stronger-torque action still gives the generated-hinge candidate 1.043 ball-diameter response error and 9.191 mm gravity peak. These are development probes, not first-attempt successes (Appendix G).

Sampling-phase and 30/60 Hz replay details are in Appendix E.

Same-input correction lineages. Table 2 traces what support and contact corrections change after an initial export. The toy image and video preserve their reconstructed object geometry across the compared XMLs. Their large reductions in contact depth show that geometry alone does not determine a usable native configuration; they do not estimate performance on unseen scenes because each final derivative was selected after inspection. The bathroom row starts from a *rigid-only fallback*: four originally predicted deformable items had already been converted to rigid bodies, so its final low penetration does not demonstrate soft-asset recovery. On the authored chest, drawer and pad scenes, top-cluster support fitting reduces contact but leaves all three required hinge/slide/volume structures absent. The same correction gives zero contact passes on ten post-outcome public applications. These paired and negative outcomes locate a specific limitation of the assembly module.

All eight additional local first attempts fail the joint gate, including export and simulation failures (Appendix Table 5).



Figure 3: Qualitative same-input comparison for the toy image: source RGB, uncorrected native assembly, and a selected support/contact derivative. Geometry comes from the same inherited backbone; the visual difference is modest even though the five-second peak contact depth falls from 35.830 to 1.690 mm. This is a selected development comparison, not a matched result from an external method.

Input scene	Targets	Fixed	Free	Hinged	Soft	$P_{\max}$ (mm)
Living room image	13/13	2	7	2	2	1.825
Toy tabletop image	11/11	1	10	0	0	1.690
Toy video masks[†]	11/11	1	10	0	0	1.929

Table 3: Selected corrected native MuJoCo development scenes shown in the teaser and supplement. "Targets" is represented/predicted, not independent scene recall. Fixed/free/hinged/soft count target assets; a hinged target can contain multiple bodies. $P_{\max}$ is the largest enabled-contact penetration across five seconds of gravity. All three trajectories completed with finite states and no numerical warnings. [†]Video masks assist object grouping, while initial geometry and pose remain largely first-frame conditioned.

4.4 First attempts and qualitative comparison

Public-image first attempts. The first three ordered public images pass in 0/3 attempts; their initial penetrations are 57.295, 57.442 and 174.812 mm. Together with eight earlier local attempts, all 11 already fail contact at initialization (Appendix D).

The frozen ten-public-RGB sequence passes in 0/10 original attempts (Appendix Table 6). Two stop upstream; eight export XML with 47/52 predicted types but only 2/7 predicted joints. Of 45 author-listed visible source targets, 43 match predicted instances; this list is not independent 3D/physics truth. Five exports fail contact initially; three start clear but later peak at 15.819, 10.927 and 3.751 mm. The last is missed at 10 Hz (saved peak 1.787 mm; Appendix Figure 6). Native existence also need not imply visual alignment: sc09 has 4/4 native predicted assets but two invisible targets, and sc12 has 8/8 but desk mask IoU 0.002. Detailed cases and an attributed external *visual-only* preview are in Appendix D.

Correction lineage. The selected scenes have unrecorded intervention budgets; their first/final XMLs and exact correction steps are in Appendix J.

Mixed scene from one image. Figure 1 shows the selected seating/tabletop scene. Its exact XML contains 13/13 predicted targets: two fixed supports, seven free rigid bodies, two hinged assemblies and two contact-enabled volumetric pillows (1,141 and 1,795 tetrahedra). Across 10,001 gravity states it has zero warnings, no collapsed/inverted elements, $P_{\max} = 1.825$ mm and minimum element-volume ratios 0.534 and 0.595. Separate force probes move both hinges and pillows (Figure 4), but do not validate true joint ranges or material parameters. Same-image predicted-mask mean IoU changes from 0.705 initially to 0.702 after gravity, masking a right-pillow drop from 0.688 to 0.562. Inpainting removes duplicated table objects without changing compiled dynamics.

Video and disturbance probes. The toy clip groups temporal masks for ten foreground objects and a tabletop; its selected model completes five seconds at $P_{\max} = 1.929$ mm, with geometry and pose still mainly frame-zero conditioned. A separate synthetic ground-acceleration probe of the

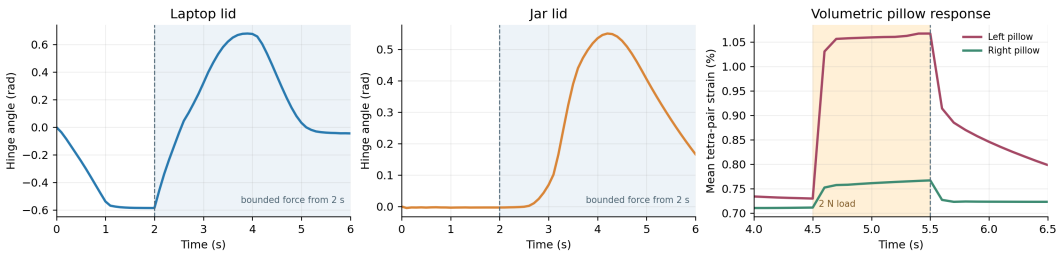

Figure 4: Physical response on the exact selected living-room XML. Separate bounded-force tests move its two native hinges and two filled tetrahedral pillows. Shaded intervals show applied probes, and the pillow curves show mean tetrahedral edge-pair strain during 2 N load and release. These motions establish executable native structure under recorded parameter priors, not recovery of true mass, elasticity or damping.



Figure 5: One tabletop arrangement, two input routes. The source frame and both five-second MuJoCo renders come from recorded image and video runs. Both scenes were selected after separate support/contact corrections; their contact peaks are 1.690 and 1.929 mm, respectively. The video branch groups temporal masks but generates geometry and pose mainly from frame zero, so this comparison does not demonstrate a temporal reconstruction advantage.

settled living scene reaches 0.546 g over 5.5 s with finite states, no warnings or inverted tetrahedra and a 1.135 mm peak. A stronger 0.772 g trial inverts a pillow tetrahedron and is rejected. Any added video camera shake/roll is a post-render transform, excluded from physical metrics.

A frozen first-frame-only run represents the same 11 predicted toy targets but fails the contact gate at 27.567 mm; the initial video-derived export also fails at 55.867 mm. The selected video derivative reaches 1.929 mm after a different assisted correction lineage. Thus this clip does not isolate a benefit from temporal input or verify video-conditioned dynamics.

5 Conclusion

SceneProbe connects image/video instance evidence to separate native simulation assets and every-step physical tests. The selected scenes demonstrate executable rigid, hinged and volumetric soft structures, while first-attempt failures expose unresolved support, mechanism and alignment errors. Contact thresholds are diagnostics rather than material or task truth: valid impacts can exceed them and wrong physical parameters can pass. The video branch still uses mostly first-frame geometry, and scene-specific correction budgets were not recorded. Matched native-output comparisons, independent action references and measured material responses are needed before claiming general physical reconstruction.

Reproducibility Statement

The anonymous supplement provides selected runnable scenes, first-attempt failures, validation records, and verification code. It omits model weights and some historical asset closures.

AI USE STATEMENT

Generative AI tools assisted with scene inventories, image-conditioned object geometry, physical-type and parameter-prior suggestions, research planning, code implementation, experiment analysis, literature search, and manuscript drafting and editing. Segmentation and inpainting used learned models. We checked reported numerical results against saved MuJoCo audits and exact XML hashes, reviewed the AI-assisted work, and take responsibility for the final text, claims and artifacts.

REFERENCES

Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, et al. Physx-omni: Unified simulation-ready physical 3d generation for rigid, deformable, and articulated objects. *arXiv preprint arXiv:2605.21572*, 2026. URL https://arxiv.org/abs/2605.21572.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *CVPR*, 2025a. URL https://arxiv.org/abs/2503.20746.

Chuhao Chen, Zhiyang Dou, Chen Wang, Yiming Huang, Anjun Chen, Qiao Feng, Jiatao Gu, and Lingjie Liu. Vid2sim: Generalizable, video-based reconstruction of appearance, geometry and physics for mesh-free simulation. *arXiv preprint arXiv:2506.06440*, 2025b. URL https://arxiv.org/abs/2506.06440.

Junhao Chen, Boran Zhang, Mingjin Chen, Henghaofan Zhang, Saining Zhang, et al. One video, one world: Turning monocular video into physical 4d scenes. *arXiv preprint arXiv:2606.31388*, 2026. URL https://arxiv.org/abs/2606.31388.

Mingyu Dong, Chong Xia, Mingyuan Jia, Weichen Lyu, Long Xu, Zheng Zhu, and Yueqi Duan. Replicateanyscene: Zero-shot video-to-3d composition via textual-visual-spatial alignment. *arXiv preprint arXiv:2604.10789*, 2026. URL https://arxiv.org/abs/2604.10789.

Yumeng He, Ying Jiang, Jiayin Lu, Yin Yang, and Chenfanfu Jiang. Spark: Sim-ready part-level articulated reconstruction with vlm knowledge. *arXiv preprint arXiv:2512.01629*, 2025. URL https://arxiv.org/abs/2512.01629.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, et al. Midi: Multi-instance diffusion for single image to 3d scene generation. In *CVPR*, 2025. URL https://arxiv.org/abs/2412.03558.

Zhao Huang, Boyang Sun, Alexandros Delitzas, Jiaqi Chen, and Marc Pollefeys. React3d: Recovering articulations for interactive physical 3d scenes. *IEEE Robotics and Automation Letters*, 11 (5):5954–5961, 2026. URL https://arxiv.org/abs/2510.11340.

Yufei Jia, Heng Zhang, Ziheng Zhang, Junzhe Wu, Mingrui Yu, et al. Gs-playground: A high-throughput photorealistic simulator for vision-informed robot learning. *arXiv preprint arXiv:2604.25459*, 2026. URL https://arxiv.org/abs/2604.25459.

Long Le, Jason Xie, William Liang, Hung-Ju Wang, Yue Yang, et al. Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. In *ICLR*, 2025. URL https://arxiv.org/abs/2410.13882.

Inhee Lee, Sangwon Baik, Sungjoo Kim, Hyeonwoo Kim, Hyunsoo Cha, and Hanbyul Joo. Simuscene: Simulation-ready compositional 3d scene reconstruction from a single image. *arXiv preprint arXiv:2606.03994*, 2026. URL https://arxiv.org/abs/2606.03994.

Haodong Li, Lulu Shao, Haolin Lu, Yu Fu, Yen-Ru Chen, Seemandhar Jain, and Manmohan Chandraker. phi-scene: Physically grounded image-to-3d scene reconstruction. *arXiv preprint arXiv:2606.21596*, 2026. URL https://arxiv.org/abs/2606.21596.

Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. *arXiv preprint arXiv:2505.18151*, 2025. URL https://arxiv.org/abs/2505.18151.

Xiaoxuan Ma, Jiashun Wang, Nicolas Ugrinovic, Yehonathan Litman, and Kris Kitani. Rest3d: Reconstructing physically stable 3d scenes from a single image. *arXiv preprint arXiv:2605.30338*, 2026. URL https://arxiv.org/abs/2605.30338.

Zhao Mandi, Yijia Weng, Dominik Bauer, and Shuran Song. Real2code: Reconstruct articulated objects via code generation. *arXiv preprint arXiv:2406.08474*, 2024. URL https://arxiv.org/abs/2406.08474.

Ben Moran, Mauro Comi, Steven Bohez, Tom Erez, Zhibin Li, and Leonard Hasenclever. Splatting physical scenes: End-to-end real-to-sim from imperfect robot data. *arXiv preprint arXiv:2506.04120*, 2025. URL https://arxiv.org/abs/2506.04120.

Nadun Ranawaka, Josiah Wong, Wei-Lin Pai, Wei-Teng Chu, Tianyuan Dai, et al. Simfoundry: Modular and automated scene generation for policy learning and evaluation. *arXiv preprint arXiv:2606.28276*, 2026. URL https://arxiv.org/abs/2606.28276.

SAM 3D Team. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. URL https://arxiv.org/abs/2511.16624.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, et al. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, 2022. URL https://arxiv.org/abs/2109.07161.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012. URL https://homes.cs.washington.edu/~todorov/papers/TodorovIROS12.pdf.

Boyuan Wang, Yue Zhang, Xutao Xue, Xueyu Song, and Yu Sun. Tableverse: A large-scale tabletop dataset with real-world grounded layouts for generalizable manipulation. *arXiv preprint arXiv:2607.21017*, 2026a. URL https://arxiv.org/abs/2607.21017.

Kaixuan Wang, Tianxing Chen, Jiawei Liu, Honghao Su, Shaolong Zhu, Minxuan Wang, Zixuan Li, Yue Chen, Huan ang Gao, Yusen Qin, Jiawei Wang, Qixuan Zhang, Lan Xu, Jingyi Yu, Yao Mu, and Ping Luo. Manitwin: Scaling data-generation-ready digital object dataset to 100k. *arXiv preprint arXiv:2603.16866*, 2026b. URL https://arxiv.org/abs/2603.16866.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, et al. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. URL https://arxiv.org/abs/2410.19115.

Ziqian Wang, Yonghao He, Licheng Yang, Wei Zou, Hongxuan Ma, Liu Liu, Wei Sui, Yuxin Guo, and Hu Su. Tabletopgen: Instance-level interactive 3d tabletop scene generation from text or single image. In *ECCV*, 2026c. URL https://arxiv.org/abs/2512.01204.

Zhuangzhe Wu, Yue Xin, Chengkai Hou, Minghao Chen, Yaoxu Lyu, Jieyu Zhang, and Shanghang Zhang. Urdf-anything+: End-to-end generation for simulation-ready articulated assets. *arXiv preprint arXiv:2603.14010*, 2026. URL https://arxiv.org/abs/2603.14010.

Chong Xia, Kai Zhu, Zizhuo Wang, Fangfu Liu, Zhizheng Zhang, and Yueqi Duan. Simrecon: Simready compositional scene reconstruction from real videos. In *CVPR*, 2026a. URL https://arxiv.org/abs/2603.02133.

Hongchi Xia, Tianhang Cheng, Wei-Chiu Ma, and Shenlong Wang. Fire3d: Feed-forward interactive 3d scene reconstruction within a minute. *arXiv preprint arXiv:2609.08848*, 2026b. URL https://arxiv.org/abs/2609.08848.

Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, et al. Cast: Component-aligned 3d scene reconstruction from an rgb image. *arXiv preprint arXiv:2502.12894*, 2025. URL https://arxiv.org/abs/2502.12894.

Xihang Yu, Rajat Talak, Lorenzo Shaikewitz, and Luca Carlone. Picasso: Holistic scene reconstruction with physics-constrained sampling. *arXiv preprint arXiv:2602.08058*, 2026. URL https://arxiv.org/abs/2602.08058.

Shujie Zhang, Jingkun Yi, Weipeng Zhong, Zirui Zhou, Yangkun Zhu, Hanqing Wang, Xudong Xu, Weinan Zhang, and Chunhua Shen. Robosnap: One-shot real-to-sim scene generation for generalizable robot learning and evaluation. *arXiv preprint arXiv:2607.06699*, 2026a. URL https://arxiv.org/abs/2607.06699.

Xin Zhang, Yabo Chen, Yijie Fang, Wanying Qu, Haibin Huang, et al. Physomni: Physics-grounded multi-object scene generation from a single image with real-time interaction. *arXiv preprint arXiv:2605.20290*, 2026b. URL https://arxiv.org/abs/2605.20290.

Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. Physsplat: Efficient physics simulation for 3d scenes via mllm-guided gaussian splatting. In *ICCV*, pp. 5242–5252, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhao_PhysSplat_Efficient_Physics_Simulation_for_3D_Scenes_via_MLLM-Guided_Gaussian_ICCV_2025_paper.html.

Shaoting Zhu, Linzhan Mou, Derun Li, Baijun Ye, Runhan Huang, and Hang Zhao. Vr-robo: A real-to-sim-to-real framework for visual robot navigation and locomotion. *IEEE Robotics and Automation Letters*, 2025. URL https://arxiv.org/abs/2502.01536.

A Detailed evaluation protocol

For scene demonstrations the unit is an *input scene* paired with a saved predicted-inventory artifact. For each predicted instance, we record whether the exact loaded XML has source geometry, visible geometry, collision representation, mass/friction where applicable, and an appropriate kinematic or soft structure. A contact-enabled cloth surface and a contact-enabled tetrahedral solid both count as a native representation for a generically predicted "deformable" target; we report their realized types separately because a cloth shell is not a filled plush model. The ratio $C_{\mathrm{pred}} = N_{\mathrm{represented}}/N_{\mathrm{predicted}}$ measures *internal completeness against a model-predicted inventory*; it cannot measure object recall against independent human annotation. In particular, a missing object in the inventory is invisible to this metric. These development inventories were revised during construction, so no immutable pre-evaluation annotation chronology is claimed.

We distinguish three further observations. (1) An input-view mask intersection-over-union compares the native rendering to automatically predicted masks from the same RGB, and is only a development alignment diagnostic. (2) Native numerical validity requires a completed five-second trajectory, finite states, no simulator warnings, and no collapsed or inverted tetrahedra. We record maximum penetration across every simulated state, rather than only a few video frames. The reported pass threshold is 2 mm. (3) A response probe checks that an applied force, joint drive or synthetic ground acceleration induces a finite, nontrivial motion in the exact model. None of these proves correspondence to unobserved real-world dynamics.

All results below use MuJoCo 3.3.7. Rigid-only scenes run at $\Delta t = 2$ ms (2,501 audited states over five seconds); the living-room model with two tetrahedral soft objects runs at $\Delta t = 0.5$ ms (10,001 states). The image cases are development inputs and include several scene-specific correction derivatives. We retain unsuccessful variants, so the rows in Table 3 are illustrative outcomes rather than a random-sample success rate. Exact source and XML hashes are in the local claim-to-evidence JSON. A convincing generalization claim would require frozen code, independent inventory/ground truth, untouched scenes, and matched recent baselines.

For an additional six-image development batch, we retained every launched cold attempt: dining, plush fruit, kitchen, bathroom, coffee table and home coffee counter. Each used seed 42, reviewed scene inventory, the same SAM 3D backend, fitted support and initial contact refinement, with no scene-specific manual override in the cold run. A pass requires all predicted targets represented in the exact XML, at least five seconds of finite warning-free gravity simulation with no invalid soft elements, and every-step peak collider penetration below 2 mm. These six images came from a local input collection; their original capture provenance is not established by our records. The target

inventories and choice of images were not independently pre-registered, so this batch is not a held-out population estimate. Its complete denominator, source-image hashes, commands, results and XML hashes are in `new_scene_attempts.json`. Subsequent corrected derivatives are excluded from the cold pass count.

Before running the current image-to-MuJoCo driver on two further multi-object local-inbox images, we fixed their image hashes, driver/configuration hashes, seed 42, SAM 3D geometry, reviewed inventory, fitted-support contact refinement and the same joint pass rule in `prospective_two_scene_protocol.json`. We include both runs regardless of export or simulation outcome. A post-review provenance check found pixel-identical images and completed reconstruction records from 22 August in this project; the frozen protocol's phrase "previously unused" is erroneous. These are predeclared runs on previously processed, selected local images, not an independent held-out test; their object inventories are model predictions. The exact correction and earlier manifest hashes are in `prospective_input_provenance_erratum.json`. Separately, we fixed one floor-pose rule and its eight-case application list in `floor_prior_repair_protocol.json` before executing those derivatives. The rule was motivated by the old development scenes, so the six-case floor results are developmental and the corrected variants never enter the cold denominator.

For a further public-input diagnostic, we selected the first three numbered RGB scene images (01–03) from the SimuScene project page before running this driver, pinned their bytes and decoded RGB hashes, and applied the same seed, review, geometry, support and five-second joint gate to all three. A scoped exact-pixel check against 144 local images and 12 dated project-session inputs found no duplicate; it cannot establish that the images were otherwise unseen. This ordered convenience sample is neither a random benchmark nor a matched numerical comparison with SimuScene. No independent target inventory, physical parameters or observed action response is provided by this protocol. The images, commands, first outcomes and all three failures are in `public_three_scene_protocol.json` and `public_three_scene_results.json`; any later repairs remain separate derivatives.

For a larger failure-inclusive public-image split, we froze ten additional numbered project-page RGBs (04–12 and 15), their hashes, a 60-entry author-written visible-asset list, the driver and runtime implementation hashes, seed 42, and the same pass rule *before* the first run. Each listed object may match at most one predicted instance by category and visible identity; ambiguous and missing matches are retained. The list excludes architecture, wall art and ceiling fixtures, and groups some visibly overlapping items. It is an author annotation, not independent all-object masks, 3D ground truth or a physical measurement. All ten source images come from a curated demonstration page, so the split is an ordered convenience sample. Official result videos for the same RGBs are pinned for source-attributed visual inspection only; they contain no competitor XML or dynamics trace for a matched simulator benchmark. The fixed inputs, annotation, execution and summaries are in `simuscene_reserved_ten_*.json`. Corrections developed after an original export are reported separately.

We additionally authored four related MuJoCo source scenes with pinned source XMLs, initial RGBs, target/type lists and prescribed actions. Each candidate used only its scene's RGB; the evaluator compared unique targets, native physical structure, source-visible masks and complete action response. Each case's target/type/action rules preceded its candidate run, but the shared evaluator implementation was frozen after the chest and drawer outcomes and before the soft outcome; we do not claim implementation-level preregistration for those earlier cases. Support corrections followed outcome inspection. One flawed drawer source and its upstream stop are retained outside the four valid-source denominator (Appendix G).

To examine the image/video difference on the toy clip, we fixed an image-driver run on its saved first-frame JPEG before execution, with seed 42 and the same SAM 3D backend as the public-image driver. The earlier video-derived model used a different assisted inventory and correction lineage. Thus the two routes can be inspected side by side, but they cannot isolate the causal effect of temporal input. The frozen settings and complete first outcome are documented in the supplemental artifact.

To test what contact checks cannot certify, we also froze a simulator-truth probe before running it: 16 combinations of primitive shape, size, mass density and sliding friction, each with a known

0.5-second settling, 0.5-second horizontal force and 0.5-second coast. For each source we evaluate the reference plus three single-factor perturbations: mass times four, friction times 0.1, or 10 mm initial embedding. We compare initial, 10 Hz and every-step contact gates with center-trajectory error under the same known force, normalized by source object diameter. This controlled synthetic probe is neither image reconstruction nor real-world physical ground truth; its full specification and source-code hash are in `known_action_audit_protocol.json`.

For a *retrospective detector comparison*, we froze a development-output snapshot of completed five-second gravity results with exactly 51 saved 10 Hz trace states and a complete per-step audit. We exclude a path if any component equals `tmp` or contains `package`, then deduplicate by XML SHA-256, keeping the lexicographically first metrics path. Other `evidence` directories remain eligible. The frozen list has 89 eligible recorded runs and 68 XMLs across 27 top-level development directories; subsequently generated attempts are outside this cohort. The accompanying `audit_cohort_manifest.json` pins all 89 source paths and file hashes. Directory count is not a count of independent source inputs. Our validator marks a run incomplete on a numerical warning, so the selected 68 are warning-free *by construction*; no independent warning-detector performance estimate follows. Many XMLs are variants of the same scene, so the 68 rows are neither independent samples nor a measure of test performance. A sparse detector tests contact penetration at the 51 saved states; the complete detector uses $P_{\max}$ over all 2,501 or 10,001 states. Both use the same 2 mm diagnostic threshold. We also subsample the already stored 10 Hz trace at 1, 2, and 5 Hz over each discrete phase, always retaining the initial state. For a targeted follow-up, we replay the 14 XMLs missed at 10 Hz, require exact reproduction of their original full-trajectory peaks, retain every per-step contact depth, and inspect 30/60 Hz nearest-step samples at all discrete phase offsets, again including time zero. This outcome-selected follow-up probes whether higher recording rates recover these particular events; it cannot estimate their miss rates on an unselected scene set. A separate local timing experiment measures audit overhead on three selected XMLs.

B Operational decision rules and implementation boundary

The three method modules name interfaces in the implemented pipeline; they do not denote newly trained perception models. The following rules identify which decisions are inherited, automatic in a specified derivative, or assisted after observing a result. The exact run manifests and script hashes determine which version generated each XML.

For accepted planar supports, the separate floor-pose diagnostic uses the fixed rule

$$z'_{\text{floor}} = \min\left(z_{\text{floor}}, \min_{i \in \mathcal{S}} z_{i,\min}^{\text{support}} - 0.75\,\text{m}\right), \tag{6}$$

where $\mathcal{S}$ is the set of accepted support fixtures and $z_{i,\min}^{\text{support}}$ is the minimum world-height of fixture i's fitted support geometry. The rule modifies neither object geometry nor contact parameters; 0.75 m is an unmeasured furniture-height prior. The top-cluster and floor rules are different derivatives with different reported denominators.

C Evidence and implementation details

The accompanying `evidence_snapshot.json` maps every image/video row and selected failure probe to its gravity metrics, SHA-256 fingerprints and verified XML bytes. The separate `audit_detector_comparison.json` contains all 68 deduplicated XMLs, their paired trace/metric hashes, time steps, sparse and complete peaks; `audit_cohort_manifest.json` pins all 89 eligible records before deduplication. The files use paths relative to the project root. The living-room gravity, jar and laptop joint probes, left and right pillow force probes, synthetic-ground acceleration, control and rejected stronger acceleration are retained as separate trajectory records in the packaged mixed scene. The anonymous archive removes machine-specific absolute paths from text metadata while preserving source and XML bytes and updating affected metadata hashes. The corrected video-conditioned toy scene is a fifth runnable package.

Implementation lineage. The reviewed inventory uses Qwen3-VL-plus for the book/ball desk and toy-video cases and Qwen3.5-plus for the new dining and plush cases, with recorded RGB prompts

Stage	Operational rule and selection status
Evidence graph	Reviewed object list, proposed type/support parent and SAM 3 masks form target nodes; SAM 3D and MoGe give per-object geometry/depth. Video masks are grouped by category, overlap and position before frame-zero lifting. *Status:* pretrained backbones with a reviewed inventory.
Support assembly	Cold fitting uses predicted support masks/depth. The later top-cluster variant takes the upper 75% of valid support heights, finds the densest 5 mm bin and keeps pixels within ± 8 mm. It requires at least 20% of mask pixels in that band, 100 valid pixels, 90th-percentile plane residual at most $\max(2\,\mathrm{mm}, 0.02 \times \mathrm{extent})$, and tilt at most 25°. *Status:* top-cluster rule selected after source-case outcomes.
Type assembly	A joint is exported only when a moving part and axis are fitted; otherwise the required joint remains absent in the audit. A closed repaired surface is tetrahedralized for filled softness when possible; a cloth shell is a separate type. The bathroom rigid fallback manually changes four proposed soft types. *Status:* selected part fits; type failures retained.
Contact/floor repair	The top-cluster derivative seats child objects with a shift cap of 0.30 predicted-object diagonal and can separate whole contact pairs. The separate floor-only derivative changes only world-floor height by Eq. (6). *Status:* post-outcome derivatives, never first-attempt replacements.
Native audit	Compile exact XML; check per-target body/type, five-second finite warning-free gravity, all-step enabled-contact depth and deformable element validity. Store extrema and sparse 10 Hz traces; regenerate full per-step traces by exact replay when needed. *Status:* frozen gate applied to each listed attempt.

Table 4: Concrete pipeline choices and selection boundary. Several knobs were chosen after observed failures; the table states their status rather than treating them as one frozen prospective algorithm.

and case-specific decoding settings; segmentations use SAM 3. A SAM 3D adapter generates per-object textured geometry with seed 42, and MoGe provides monocular geometry cues. A predicted support mask and scene depth are converted to a finite support surface; LaMa fills RGB areas hidden by the foreground before that image is used as a support texture. Rigid collision shapes use convex decomposition. Soft meshes are tetrahedralized from a repaired generated surface, with a reported rigid fallback if meshing fails. These components are not introduced as new learned models. The toy-video branch samples the 28.67-second original clip to three frames per second; temporal tracks are grouped before per-object SAM 3D generation at frame zero. The file-level manifests record the model checkout revisions, prompts, masks and per-stage commands.

Physical priors in the exact living-room model. The simulator masses are 0.532 and 1.867 kg for the left and right soft pillows, 1.500 kg for the laptop assembly, 0.500 kg for the jar assembly, and 0.191 kg for the pumpkin. The soft pillow sliding-friction prior is 0.7; jar/laptop sliding friction is 0.6. The jar joint range is $[0, 2.2]$ rad and the laptop range is $[-0.6, 1.0]$ rad. The moving laptop screen mesh was selected using an input-mask-guided local plane at $z = -100$ mm, with 1.2/0.3 kg body masses. These values are operational model choices, not measurements from the image. The living XML disables flex self-collision; the all-step audit checks enabled simulator contacts and individual tetrahedra, not nonlocal visual-surface intersection.

Local audit cost. We ran three pinned XMLs for five simulated seconds without rendering, twice per condition in reversed order: raw `mj_step`, `mj_step` plus `mj_forward` at each state, and the latter plus `StepAudit.observe`. Median wall times (s) were respectively 134.097/272.863/295.097 for the volumetric living room, 3.028/5.700/6.168 for rigid toys, and 27.139/53.686/59.916 for the outdoor surface-soft scene. The audit itself added 8.1%, 8.2% and 11.6% over the per-state forward loop; the full checked loop took 2.20, 2.04 and 2.21 times raw stepping. All three modes reached identical terminal `qpos` bytes within each scene. These local timings exclude XML compilation and rendering, overlap other machine workloads and do not gen-

eralize across hardware. Exact trials, code and hashes are in `audit_cost_results.json`; the earlier one-second timing diagnostic is retained separately in `audit_overhead.json`.

Cached-state audit diagnostic. For the tested `implicitfast` XMLs, MuJoCo's `mj_step` leaves contact and flex-position arrays describing the pre-integration state. A diagnostic wrapper audits those cached arrays at the previous time and calls `mj_forward` once at the endpoint. On four exact XMLs (living, toys, outdoor, and the toaster's early tetrahedron inversion), it matches the existing every-step audit's completion, checked-state count, simulated time, warning vector, peak contact depth and first invalid soft state. The three completed scenes also match terminal `qpos` bytes; the unstable toaster trajectory does not, despite detecting the same inversion at 0.014 s. A different forward-evaluation schedule can alter unstable dynamics, so this variant cannot replace the conservative validator without wider equivalence testing. Single local loop times exclude compilation and rendering and are not a paired speedup estimate. Exact diagnostic protocol, source and outputs are included separately.

Read-only callback audit diagnostic. To avoid changing the integration schedule in that unstable case, we also ran a post-outcome audit from MuJoCo's control callback, observing the pre-integration contact and flex state without calling `mj_forward` inside the loop. Across the exact toy, outdoor, toaster and living XMLs, the complete `qpos`/`qvel` byte stream matches a raw `mj_step` run for the same number of steps; the four callback reports also match their saved conservative gravity audits on completion, checked-state count, peak penetration and first invalid soft state. The toaster still stops at the first tetrahedron inversion at 0.014 s after eight checked states; its seven integrated steps match raw stepping exactly. The control callback is process-global and would have to be composed with a pre-existing controller. Exact XMLs, traces, reference hashes and source are in `callback_step_audit_results.json`.

A paired post-outcome local timing probe ran raw stepping, callback auditing and per-step-forward auditing in one order and its reverse on the three completed XMLs. Median five-second wall times were 2.920/3.348/6.374 s (toys), 31.402/32.861/62.550 s (outdoor) and 145.541/157.420/300.779 s (living), in that mode order. The callback therefore cost 1.05–1.15 times raw stepping in these trials, while the forward audit cost 1.90–1.91 times the callback. All modes had identical terminal `qpos`/`qvel` hashes and the two audits had identical checked-state counts and contact peaks. The timing excludes compilation and rendering, overlaps other workloads and is not a hardware-general speed claim. Both ordered trials, the frozen protocol and source are in `callback_cost_results.json`. These results concern the three tested `implicitfast` gravity models, not controlled actions or other integrators.

Failure and comparison scope. The 68-XML detector corpus includes same-scene variants and corrected derivatives; no independent-scene uncertainty interval is implied by its counts. The source inventory for the toy-video branch originally had no native table identifier. A derived, hash-linked audit mapping pairs that predicted support target with the later `observed_support` body. Auditing the original unmapped inventory gives 10/11, and auditing the derived native mapping gives 11/11. This mapping changes neither the source inventory nor the MuJoCo XML. The more challenging plush scene failed as both a cloth-shell and a volume-soft representation. The recorded trajectory identifies failure but does not isolate its cause.

D PUBLIC-IMAGE FIRST-ATTEMPT SEQUENCE

The first three ordered SimuScene project-page RGBs were run through the same image driver as original first attempts. Scene 01 represents 6/6 predicted targets (two fixed rigid, three free rigid, one surface soft) but stops numerically at 0.007 s; scene 02 represents 6/6 predicted rigid targets and completes five seconds without warnings, yet has 57.442 mm peak contact penetration; scene 03 represents 6/7 targets (one fixed rigid, one free rigid, four surface soft, one missing remote) and stops at 0.0015 s after contact proliferation and a MuJoCo solver-stack overflow. Their initial maximum depths are 57.295, 57.442 and 174.812 mm. Scene 01's interrupted peak is not a comparable five-second maximum. Scene 02's input-view render omits several tabletop items despite 6/6 internal representation, illustrating why predicted-inventory coverage is not visual recall. The exact first-

New RGB input	Targets	Time (s)	$P_{\max}$ (mm)	Outcome
Dining tabletop	11/11	5.000	34.727	Contact gate
Plush fruit basket	12/12	0.0005	–	Numerical stop
Kitchen counter	11/12	0.008	–	Tetrahedron inversion
Bathroom counter	11/11	0.002	–	Numerical stop
Coffee table	11/11	5.000	18.462	Contact gate
Home coffee counter	8/11	5.000	17.755	Coverage/contact
Toaster tabletop†	7/7	0.014	–	Tetrahedron inversion
Outdoor tabletop†	9/10	5.000	25.988	Coverage/contact

Table 5: All eight additional whole-image first attempts of this driver. The first six are development cases; †the last two driver/configuration choices were fixed before execution on images previously processed in this project. They share seed 42, SAM 3D and fitted-support contact refinement. Dining requested default thinking/90 s, then its failed review attempt was resumed with thinking off/240 s; the other seven used off/240 s. "Targets" compares native representations to each predicted inventory, not human object recall. Contact-enabled surface cloth counts as a generic deformable representation but is distinct from a filled solid. $P_{\max}$ is omitted for interrupted trajectories because their terminal contacts are not a meaningful five-second comparison. None passes the joint full-coverage, five-second finite warning-free and under-2 mm contact gate. The six first run commands, input/XML hashes and failure stages are in `new_scene_attempts.json`; the latter two are in `prospective_two_scene_protocol.json` and `prospective_two_scene_results.json`.

export XMLs, source images, hashes, native audits and failure records are retained in the archive. These public source images are not a matched SimuScene output comparison.

For the separate ten-image sequence, the source RGB hashes, ordered IDs, author-written 60-entry visible-object list and frozen driver hashes precede the first run. Table 6 retains every upstream and native failure. Sc07's reviewed inventory is rejected by a support-parent regex that misreads "on floor left of desk" as "on desk"; sc11's first vision response contains JSON line comments and fails strict parsing. Neither produced a native first-export XML. These are pipeline failures, not excluded samples. Five exported scenes violate the contact target already at $t = 0$; the other three first cross it during gravity. The two interrupted numerical trajectories have no comparable five-second contact peak. The source-image list deliberately excludes architecture and some fixtures, and after execution we discovered it missed one of sc12's five visible plant pots. We retained the original list and counted that predicted pot only as an extra prediction, so even 7/7 listed matches there is not complete scene recall. Source-image matching is author adjudication, not external annotation. The per-object simulator type, mass and sliding friction for all 52 predicted assets in the eight native XMLs are recorded in `simuscene_reserved_physical_properties.json`; none is a real-action measurement.

Post-outcome parser recovery. After freezing all ten original outcomes, we cloned the sc07 and sc11 failed directories and patched only JSON comment handling and the overbroad tabletop-support phrase test. Sc07 reuses the exact saved discovery and review model texts; it then completes segmentation and geometry but stops at native export because a predicted deformable plant has no usable material prior. Sc11 reuses the exact saved first vision response, obtains a new reviewed inventory, and reaches a five-second native rollout with zero warnings; however, its predicted articulated desk has no realized joint and the initial/maximum contact depth is 57.441 mm. The separate derivation and exact stages are in `simuscene_reserved_posthoc_upstream_recovery_results.json`. Neither derivative alters the frozen 0/10 first-attempt denominator or establishes an end-to-end success.

Post-outcome support and floor composition. We applied the previously fixed 0.75 m floor-gap rule to all ten public cases after their top-horizontal support and seating outcomes were known. Five prior stage stops remain stops. Among the five completed derivatives, the maximum contact depths change from 29.898, 6.171, 9.501, 4.369 and 5.136 mm to 39.718, 6.171, 26.068, 3.218 and 5.274 mm for sc04, sc09, sc10, sc12 and sc15. Thus the composition gives 0/10 contact passes. A separate 0.25 m free-body retention gate also gives 0/10 joint passes: the largest

Public RGB	Visible match	Native/pred.	Time (s)	P_0 (mm)	$P_{\max}$ (mm)	Original outcome
sc04	4/4	5/5	5.000	0.000	15.819	Contact gate
sc05	6/7	7/9	0.272	22.023	–	Numerical stop
sc06	7/8	6/8	0.022	117.018	–	Invalid state
sc07	–	–	–	–	–	Inventory stop
sc08	6/6	6/6	5.000	49.403	49.403	Contact gate
sc09	3/3	4/4	5.000	140.806	140.806	Contact gate
sc10	6/6	6/7	5.000	0.000	10.927	Coverage/contact
sc11	–	–	–	–	–	Inventory stop
sc12	7/7	8/8	5.000	0.000	3.751	Contact gate
sc15	4/4	5/5	5.000	7.675	7.675	Contact gate

Table 6: All ten ordered public-image first attempts of the frozen driver. Visible match is one-to-one correspondence between the author's pre-run source-image list and predicted instances, evaluated only when a native inventory exists; it is not independent 3D/physics ground truth. Native/pred. counts type-consistent native representations against the model's own inventory. P_0 is initial maximum enabled-collider depth; $P_{\max}$ is reported only for completed five-second trajectories. All upstream failures are retained. No row is a matched numerical SimuScene comparison.

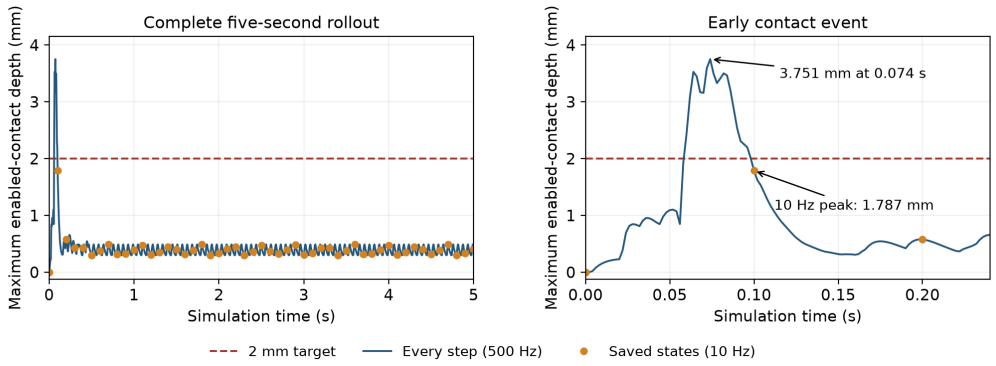

Figure 6: Exact first-export sc12 MuJoCo replay: enabled-contact depth at all 2 ms states versus the states retained at 10 Hz. The first 0.24 s is enlarged at right. The 3.751 mm event at 0.074 s crosses the development target between saved frames; the 10 Hz maximum is 1.787 mm. This is one observed contact event, not evidence of action failure or a population miss rate. The exact XML hash, replay script and every-step values are supplied in the artifact.

final displacements in the five completed cases are 3.032, 0.426, 0.604, 0.015 and 0.379 m. Moving the floor can remove an initial collision while allowing unsupported objects to fall. The rule was selected after inspecting these images and is a negative development diagnostic, not a held-out improvement or full-scene reconstruction. Frozen inputs, XML hashes, five-second metrics with every-step contact summaries, saved sparse traces, final states and alignment records accompany `reserved_support_floor_composition_posthoc_results.json`; individual every-step states are regenerated by replay rather than stored.

E AUDIT CADENCE SENSITIVITY

Subsampling the saved 10 Hz trace at 1, 2, and 5 Hz detects respectively 14–21, 14–21, and 21–22 of the same 38 complete-audit violations across discrete sampling phases; 10 Hz detects 24. Each rate also checks the initial state, using respectively 6, 11, 26, and 51 stored states per trajectory. This is a post-hoc sampling sensitivity analysis of the same data, not an independent evaluation or a measurement of audit runtime.

We also replayed all 14 exact XMLs whose violation the saved 10 Hz phase missed; their every-step and 10 Hz peaks matched the original records exactly. At phase zero, nearest-step 30 and 60 Hz inspection detects 8/14 and 10/14. Across all discrete phase offsets, only 1/14 is detected at every 30 Hz phase and 5/14 at every 60 Hz phase, although each violation is detected at some phase of both rates. The conventional median longest contiguous interval above the 2 mm threshold is 9 ms.

These 14 XMLs were selected *because* 10 Hz missed them and include related development variants; the figures characterize those transient events, not new-scene miss probabilities. The exact per-step depths and sampling code are in `replay_missed_cadence.json`.

F KNOWN-ACTION SYNTHETIC GATE DIAGNOSTIC

The prespecified source grid has two primitive shapes (sphere and cube), radii/half-extents 0.05 and 0.10 m, sliding friction 0.3 and 0.8, and density 500 and 1000 kg/m^3: 16 source models. The nominal mass is geometry volume times density. A constant horizontal force equal to $1.5 m_{\text{source}} g$ acts during the 0.5–1.0 s interval, after 0.5 s of settling; the body then coasts to 1.5 s. MuJoCo 3.3.7 uses 2 ms steps. Source and perturbed center positions are paired at 10 Hz during force and coast. We declare response error when the maximum aligned Euclidean center-position difference exceeds 0.1 source diameter; this arbitrary operational threshold is not a real task-success criterion. The two unperturbed references with every-step contact excursions are the 0.05 m half-extent, 0.8-friction cubes at both densities: they reach 8.874 mm between 10 Hz states, while their 10 Hz peaks are 0.250 mm. The full trajectories, source code hash, XML generation rule and alternative 0.05/0.25-diameter thresholds are in `known_action_audit_protocol.json` and `known_action_audit_results.json`.

Variant (16 sources each)	Initial fail	10 Hz fail	Every-step fail	Response fail
Unperturbed reference	0	0	2	0
Mass $\times 4$	0	0	0	16
Friction $\times 0.1$	0	0	0	16
Initial embed 10 mm	16	16	16	0

Table 7: A controlled, dependent perturbation study of a contact-depth gate (2 mm) versus a prespecified trajectory-error gate (0.1 source diameter). Every case completes with finite states and zero MuJoCo warnings. The parameter perturbations preserve initial contact geometry; the embedded object settles before the force probe. Thus the mismatch between contact and response outcomes is by design, and does not estimate failure rates for reconstructed scenes.

The same 32 parameter-perturbed models exceed both the 0.05- and 0.25-diameter response thresholds, while none of the 16 embedded variants does. These threshold checks use the same trajectories; they are sensitivity analysis, not extra trials.

After inspecting the 2 ms outcomes, we repeated the full 16-source, four-variant grid at 1 and 0.5 ms with the same real-time force schedule and 10 Hz observation grid. At both finer steps, the 32 mass/friction variants still have zero 2 mm contact-gate failures and 32/32 response failures, while all 16 embedded variants violate contact and none violates the response gate. Unperturbed reference every-step contact violations rise from 2/16 at 2 ms to 4/16 at both finer steps; the newly counted pair is the 0.10 m, 0.8-friction cube at both densities. This is an explicitly post-outcome numerical sensitivity analysis of dependent synthetic cases, not evidence that smaller steps systematically worsen physical fidelity. The frozen extension and all per-scene results are in `known_action_timestep_protocol.json` and `known_action_timestep_results.json`.

G TRUTH-RENDER-RECONSTRUCT-ACT PILOT

We authored one synthetic MuJoCo 3.3.7 source scene with a fixed table, free red ball, free blue box and hinged laptop. Its 640×480 RGB render is the only reconstruction input; source XML, true instance masks, masses and the held-out action trace are evaluator-side information. The source scene/script hash, four-target denominator and source action were frozen in `truth_render_scene_pilot_protocol.json`. Unique category-matching rules, type checks and the full action metric were frozen in `truth_render_reconstruction_eval_protocol.json` before candidate reconstruction began. The run used GPU 2 instead of the planned GPU 6 while retaining the same input, seed, reviewed inventory, SAM 3D geometry, support refinement and five-second gravity export; this

pre-outcome hardware deviation is recorded in `truth_render_pilot_execution_note.md`. The execution note also pins the parser-recovery discover/review code hashes used before this pilot, which differ from the earlier ten-image protocol; this pilot uses a separate input and implementation revision and does not extend that frozen denominator. This is one authored scene, not natural-image or external-system evidence.

The review pass predicts all four objects and a revolute laptop joint. The native XML contains all four body identities but no laptop descendant hinge, so the prespecified complete-scene action test fails by definition. Its five-second gravity trace is finite and warning-free but begins with 73.402 mm penetration. Comparing source-XML instance masks to native input-view masks without fitting to source labels gives IoU 0.380 for the table, 0.793 for the ball, 0.656 for the box and 0.000 for the laptop (macro 0.457). In the same camera the laptop has only 117 native visible pixels against 6,255 source pixels. Compiled candidate/source total-mass ratios for ball, box and laptop are 4.82, 6.18 and 3.03; the source mass is hidden from the single-RGB predictor, so these are diagnostic errors rather than proof that another image-only method can recover the correct values. The exact XML hashes, masks and mass arithmetic are retained in the supplemental source-truth evaluation records.

The source action settles for 0.5 s, applies 0.45 N to the ball and 0.055 Nm to the laptop hinge for 0.5 s, and coasts for 0.5 s. A replay of the unchanged source reaches 3.628 mm penetration for 19 discrete states during the ball–box impact at 0.968 s; its longest continuous above-2 mm interval is 38 ms, its saved 10 Hz peak is 1.371 mm, and it has zero warnings. Thus the arbitrary 2 mm gate rejects even a geometrically exact source model during intentional contact. This is a simulator-truth counterexample, not a real-world material measurement. Since the reconstructed laptop hinge is missing, we separately ran a *post-outcome rigid-only* 0.45 N ball probe, retaining the failed full-scene outcome: its maximum image-right trajectory error is 0.413 source-ball diameters and its terminal error is 0.062 diameters. The partial probe omits the source's simultaneous hinge actuation and cannot serve as a complete-scene action score. The full source replay and partial diagnostic are in `truth_render_source_audit.json` and `truth_render_rigid_action_posthoc.json`.

Related source-truth suite and accounting. We later authored a hinged chest, a cabinet with a sliding drawer, and a table with a volumetric soft pad. Each source XML, initial RGB, source action and evaluator-side target/type list was pinned before its separate candidate run; the same image/SAM 3D/support driver was used, with only RGB given to reconstruction. The chest lid is weakly visible in the source image, so its missing hinge is a difficult single-view case. An earlier drawer source had 4 mm of initial shell/drawer self-overlap and its candidate stopped during inventory review; it is retained as a source-design negative control and excluded from the four valid-source denominator. We replaced the solid shell with cavity panels and froze the valid drawer v3 before running its candidate. All four valid source scenes begin with zero contact penetration and are warning-free under their prescribed action. Their impact peaks of 3.252–6.011 mm show why the 2 mm development target cannot be treated as action correctness.

RGB source	Unique targets / source	Required type recovered?	Source-mask IoU macro	Original $P_{\max}$ (mm)	Corrected $P_{\max}$ (mm)	Full action pass?
Table/laptop	4/4	no hinge	0.457	73.402	9.191†	no
Chest	2/4	no hinge	0.338	117.805	1.543	no
Drawer v3	2/3	no slide	0.435	71.728	0.991	no
Soft pad	2/4	no volume	0.276	82.372	0.612	no

Table 8: Four related, authored RGB-only source reconstruction tests. Unique target matching and source-mask IoU are for each *original first candidate*; ambiguous identities count as unmatched. All four original gravity runs finish five seconds with no numerical warnings. The corrected column is explicitly post-outcome development and retains original failures. †The table/laptop corrected value uses a later top-horizontal, seated and generated-hinge combination; unlike the other three contact-only corrections, it adds a hinge but still fails contact and source action.

For the chest, drawer and soft scenes, the same top-horizontal/seating rule accepts predicted support pixels and reduces the five-second contact peak below 2 mm. The relevant missing hinge, slide and volumetric elements remain absent, so none passes the prescribed full action test. Input-view *predicted-mask* IoU changes from 0.685 to 0.713, 0.649 to 0.606 and 0.322 to 0.734, respectively;

drawer alignment worsens. These are same-image predicted-mask comparisons, whereas the source-mask IoU in Table 8 uses evaluator-side authored XML masks. The top-horizontal rule was chosen after inspecting the original table/laptop scene, and all three first outcomes were already known before its batch application. In the ten previously failed public RGB cases, an identical post-outcome rule gives five completed corrected trajectories with peaks 29.898, 6.171, 9.501, 4.369 and 5.136 mm for sc04, sc09, sc10, sc12 and sc15. Sc05, sc06 and sc08 stop at seating, fit or accepted-support selection; sc07 and sc11 remain original upstream stops. The contact pass count is 0/10. The frozen public first-attempt 0/10 full-scene result is unchanged.

Post-review mask-association diagnostic. The frozen unique category rules assign no blue-cube target when both "cube" and "box" are predicted in the chest case; the soft case also has two predicted cubes. This conflates semantic association with visible geometric overlap. We therefore froze a separate evaluator-side one-to-one assignment that maximizes source/native mask IoU over all predicted native instances, irrespective of category, after all first outcomes were known. The original category-matched macro IoUs 0.338, 0.435 and 0.276 become oracle-assigned upper bounds 0.688, 0.653 and 0.320 for chest, drawer and soft. In chest, oracle IoU is 0.745 for the blue cube and 0.655 for the chest-shaped box; in the soft case it is only 0.036 for the red pad. These are geometric diagnostics using hidden source masks, *not* image-only inferred identities or revised first-attempt scores. Joint/slide/flex and full-action failures remain unchanged. The deterministic assignment matrix, frozen XML/input hashes and original scores are in `posthoc_mask_assignment_results.json`.

Post-review source hinge-unit erratum. The original laptop and chest XMLs specified joint ranges `-0.9 0.9` and `-0.8 0.8` without setting an XML angle unit. MuJoCo compiled these as ± 0.015708 and ± 0.013963 rad, i.e., $\pm 0.9°$ and $\pm 0.8°$, rather than the intended ± 0.9 and ± 0.8 rad. This weakens the original hinge-action benchmark. We preserve the frozen source XMLs, candidate reconstructions and round-6 reviews, and provide separate corrected source XMLs with explicit degree-valued ranges that compile to the intended radians. The exact initial compiled body/geom poses and the reconstruction-input PNG bytes are unchanged. Independent EGL re-renders differ in fewer than 0.1% of pixels by at most one 8-bit channel value; an earlier pixel-exact protocol stopped before action replay and is retained. Under the original 0.055/0.05 Nm torques, the corrected laptop/chest hinges still change by only 0.0034/0.00025 rad after the 0.5 s settle, so the originally reported 0.007 rad incremental hinge error does not validate mechanism response. Re-comparing the unchanged generated-hinge candidate to the corrected original-torque laptop source reduces its full-trajectory hinge-change magnitude error from 1.564 to 0.677 rad, but the ball error stays 1.043 diameters. All values, trace hashes and unit checks are in `source_hinge_angle_posthoc_v2_results.json`.

A source-only torque sweep selected stronger, explicitly post-outcome interventions. At 0.22 Nm, the corrected laptop source has 1.882 rad driven hinge motion; at 5 Nm, the corrected chest source has 0.926 rad, both warning-free. With the same stronger laptop action, the generated-hinge candidate completes 1.5 s but has 1.303 rad maximum hinge-change *magnitude* error, 1.043 source-ball diameters maximum response error and 9.191 mm contact peak. A post-review audit found that the earlier 0.633 rad "driven incremental" number subtracted absolute changes from the initial pose, losing source motion when the hinge crossed zero. Recomputing the absolute difference of each trajectory's displacement *from its own 0.5 s settled pose* gives 1.876 rad maximum error at 0.8 s; the source moves 1.882 rad and the candidate only 0.0074 rad from those poses. The old and corrected formulas and exact unchanged samples are retained in `postsettle_hinge_metric_results.json`. The chest candidate has no hinge and fails before response scoring. The large chest source torque and magnitude-only angle scores are synthetic probes, not calibrated real mechanics or pose fidelity; no cross-model joint-axis or part-pose registration is claimed. The source-only selection and paired post-outcome result are in `corrected_source_hinge_torque_sweep_results.json` and `posthoc_substantial_hinge_action_results.json`. The anonymous artifact's source-suite manifest links the unchanged first candidates, corrected sources and associated records.

Controller composition diagnostic. Using the same authored source XML and simultaneous force/torque schedule, we composed a pre-existing control callback with read-only contact auditing. For `implicitfast`, Euler and RK4, the complete integrated `qpos`/`qvel` byte streams, terminal hashes, elapsed times, warnings and controller call counts exactly match the controller-

only runs. The callback observes 750 samples under the first two integrators and 3,000 RK4 intermediate samples for 750 integrated steps; those samples do not include every final integrated endpoint. A separately frozen post-review endpoint audit checks the final 1.5 s state in each run after safely removing/restoring the controller callback and confirms the raw/composed integrated streams are unchanged. This corrects an overbroad earlier callback summary label; it is a scoped implementation result, not a proof for arbitrary controllers or continuous trajectories. Protocols and exact hashes are in `callback_control_composition_results.json` and `callback_endpoint_posthoc_results.json`.

Post-review public-scene contact ablation. Before rerunning the three already observed public scenes, we froze one shared switch that adds whole-pair separating constraints to the existing initial contact refiner, retaining their same reconstructed meshes and inventories. The optimizer made zero translation in all three cases. All still fail the joint gate; their initial penetrations remain 57.295, 57.442 and 174.812 mm. Scene 02 still completes five seconds with a 57.442 mm peak, while scenes 01 and 03 stop early. The same-input predicted-mask mean IoU does not change because the export poses do not change. This negative development ablation, including failed optimizer status, is in `public_three_pair_separation_dev_results.json`; it cannot be counted as three additional first attempts or as a successful automatic correction.

Depth-support development probe. A different post-review probe retains rendered furniture meshes but replaces their original fixed collision hulls with one or more horizontal slabs inferred from predicted depth near each child object. A common version on the same three observed public scenes passes the physical gate in 0/3: scene 01 still stops after 0.001 s, scene 03 after 0.002 s, and scene 02 completes five seconds but peaks at 41.923 mm (from 57.442 mm). An additional scene-02-only horizontal pair-separation derivative has zero initial penetration and a 2.375 mm five-second peak, still outside the strict gate; it shifts the two books by 111 and 123 mm and lowers same-image predicted-mask mean IoU from 0.290 to 0.219. This probe disables the original furniture collision hulls, so it cannot certify side contacts or hidden support structure. Its reduced contact peak is neither a visual improvement nor an independently validated reconstruction. Exact XMLs, settings, alignments and failures are in `public_three_support_patch_dev_results.json`.

G.1 Complete fixed floor-pose ablation

We applied the fixed floor-pose rule in Eq. (6) to all six failed cold image exports, without moving any object or changing its mass, shape or contact parameters. The same-input predicted-mask mean IoU rose in all six cases (mean change $+0.063$, range $+0.039$ to $+0.096$), largely because the infinite floor stopped occluding the fitted tabletop (Figure 7). None passed the joint gate. Dining and coffee still completed five seconds with unchanged 34.727 and 18.462 mm peak penetrations. On the two additional inputs, toaster IoU rose from 0.585 to 0.712 without preventing its 0.014 s tetrahedron inversion; outdoor IoU rose from 0.688 to 0.733 while its five-second peak remained 25.988 mm and the tape-measure articulation remained missing. Table 9 includes all eight cases declared for the same 0.75 m floor-height rule. The changed XMLs and alignment/audit records are pinned in `floor_prior_repair_results.json`; no case was removed for a bad result. The rule was devised after inspecting development scenes, and the masks are predictions on the same images. Interrupted soft runs have no comparable five-second $P_{\max}$, and a filled-plush interpretation is not supplied by the fruit scene's nine cloth shells.

Public sc09 floor-only post-outcome probe. The original fixed support candidate for sc09 was rejected by the automatic planar-support report. After seeing that the original native view hides the pen holder and eraser, we manually selected that predicted support for the previously used 0.75 m floor-gap rule, changing only the infinite floor height from 0 to -0.787 m. Initial collider penetration falls from 140.806 to 0 mm, but the derived five-second trajectory still reaches 41.553 mm and the holder and eraser remain at zero visible pixels; same-image predicted-mask mean IoU remains 0.370. This is a negative same-scene diagnostic and is excluded from first-attempt counts. The exact XMLs, runs and image-mask records are in `sc09_floor_posthoc_results.json`.

Input	Old IoU	Floor IoU	Floor time (s)	Floor $P_{\max}$ (mm)
Dining tabletop	0.643	0.701	5.000	34.727
Plush fruit	0.584	0.623	0.0005	–
Kitchen counter	0.493	0.536	0.002	–
Bathroom counter	0.506	0.582	0.002	–
Coffee table	0.675	0.738	5.000	18.462
Home coffee	0.505	0.602	5.000	17.755
Toaster tabletop	0.585	0.712	0.014	–
Outdoor tabletop	0.688	0.733	5.000	25.988

Table 9: Same-input predicted-mask IoU before/after moving only the infinite floor. The first six inputs are development cases; the last two images and the floor rule were fixed before their respective first repair runs. $P_{\max}$ is shown only for completed five-second trajectories. None of the eight derivatives meets the joint pass rule.



Figure 7: Same input view and identical object geometry with the original $z = 0$ infinite floor (middle) versus the fixed floor-height derivative (right). The rule restores the visible finite table but neither scene meets the 2 mm contact gate. Panels are cropped/resized exact renders; source hashes are recorded with the figure.

H POST-REVIEW SOURCE-PARAMETER CONTROLS

We froze two deterministic control protocols before running their listed parameter variants, after the underlying four authored scenes and force schedules had been developed. These controls ask whether a gravity contact check predicts response when source-model geometry is exactly preserved; they are not image/video reconstruction results and the parameter ranges are intentionally severe. Each model uses its original RGB, XML geometry, joint topology and prescribed 1.5 s action. All outcomes are simulator-defined, with no measured real-world mass, damping or material reference.

The first protocol multiplies the moving-ball mass and inertia in each of four source scenes by 0.25, 0.5, 1, 2 or 4. The original implementation changed compiled `body_mass` and `body_inertia` without calling MuJoCo's `mj_setConst` to refresh derived inertia and inverse-weight fields. Its reported 20/20 response agreement is invalid and retained only as a versioned error record in `controlled_parameter_audit_results.json`. In the post-review correction, we changed no cases, forces, factors or thresholds, called `mj_setConst`, and matched six compiled mass/inertia-related fields to fresh compilation of each mass-edited source XML. At the original maximum ball-trajectory tolerance of 0.25 source-ball diameters, the four unedited references agree and all 16 perturbed variants fail. The 10 Hz five-second gravity contact check passes all 20, so it misses all 16 response failures. The every-step check passes 15: relative to response agreement it has 3 true passes, 12 false passes, 1 false rejection and 4 true rejections. Its five soft-scene rejections reflect a 2.930 mm transient peak missed at 10 Hz (1.413 mm). The corrected intentional-action con-

tact peaks range from 0.369 to 11.885 mm, and even exact references can exceed the 2 mm contact threshold during an intended action. Per-variant corrected errors, contact summaries and old-to-new label changes are in `controlled_parameter_audit_setconst_v2_results.json`. This is an explicit post-review correction of dependent authored source controls, not an independent reconstruction test.

The second protocol changes only the compiled damping of the laptop hinge, chest hinge or drawer slide by factors 0.01, 1, 100 or 1000. All 12 five-second gravity runs finish warning-free and below 2 mm at every step. We compare each variant's *signed* joint displacement from its own 0.5 s settled state with its unedited same-source trajectory, using maximum 10 Hz error and thresholds fixed at 0.25 rad for hinges and 0.02 m for the slide. Eight variants fail response agreement while passing contact: three laptop, three chest and two drawer. At $1000\times$ damping, the laptop's maximum error is 1.879 rad and the drawer's is 0.160 m; the respective gravity contact peaks are only 0.551 and 0.369 mm. The low-damping laptop/chest variants have large transient trajectory errors even though their 1.0 s progress nearly matches the reference. Thus 8/12 is specific to a maximum-trajectory criterion, not terminal task failure or general real-physics accuracy. The fixed protocol, all 12 rows, current-interpreter replay and exact source hashes are in `controlled_joint_damping_audit_results.json`. These counterexamples motivate an independently validated response/task gate, but do not supply one for reconstructed scenes.

I Additional selected scene demonstrations

These are selected development derivatives, not samples from the frozen first-attempt denominators. The gallery contains three distinct source arrangements; the image and video toy rows use the same scene content and count as two input routes, not two independent environments. Each panel shows source RGB and native MuJoCo output from recorded runs. The video panel is an initial scene assembled with temporal mask grouping, not a dynamic reconstruction of the whole clip.

I.1 Coverage of public video scene families

We also inspected the nine scene families shown in SimFoundry's video gallery (Ranawaka et al., 2026). Existing records for eight families use a single image from the source family; they were produced under differing development configurations. The living-room row is a selected, assisted image derivative. These entries are an input-coverage audit, not nine full-video reconstructions or a matched comparison with SimFoundry's native outputs. Table 11 includes failures rather than selecting only the successful gallery rows.

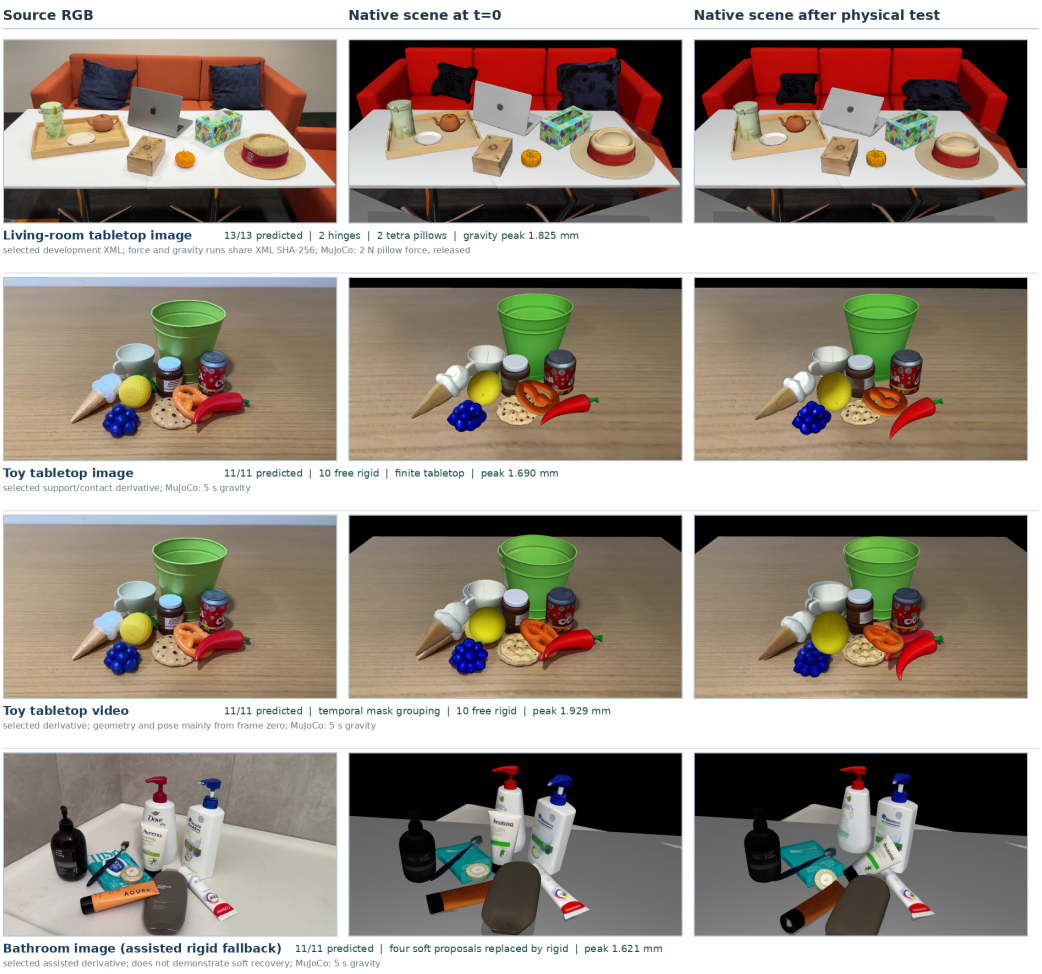

Figure 8: Selected image/video scenes with exact source RGB, native initial render and later native response frames. The living-room row shows a 2 N force probe applied to a tetrahedral pillow after gravity settling; its model also has two hinges. The other rows show five-second gravity rollouts. The bathroom is an assisted rigid fallback: four proposed soft types were replaced, so it supplies no evidence of soft reconstruction. Counts are relative to predicted inventories, and numerical values are from exact XML audits; material constants remain priors. The toy-video initial mesh and pose are mainly conditioned on frame zero.

Native component	Realized structure	Stored parameters (prior)	Executed check
Laptop lid	hinge, range $[-0.6, 1.0]$ rad	damping 0.02, friction loss 0.4	bounded joint-force motion
Jar lid	hinge, range $[0, 2.2]$ rad	damping 0.01, friction loss 0.05	bounded joint-force motion
Left pillow	1,141 tetrahedra, 0.532 kg	$E = 3$ kPa, $\nu = 0.3$, $\mu = 0.7$	$J_{\min} = 0.534$; 2 N probe
Right pillow	1,795 tetrahedra, 1.867 kg	$E = 3$ kPa, $\nu = 0.3$, $\mu = 0.7$	$J_{\min} = 0.595$; 2 N probe

Table 10: Physical properties of the selected living-room XML: inspectable native topology and numerical simulator parameters with executed motion/element checks. E is Young's modulus, ν Poisson's ratio, μ sliding friction, and $J_{\min}$ the minimum signed tetrahedral-volume ratio during five seconds of gravity. Masses, ranges and material constants are inferred or chosen priors; no ground-truth material identification is claimed. The left-pillow force probe reaches 0.375% peak pair strain under this particular load and retains 0.112% after release.

Source family	Image-route scope	Simulated (s)	$P_{\max}$ (mm)	Status
Bathroom	existing export	0.002	650,152.116	halted
Coffee	existing export	5.000	18.462	contact fail
Dining	existing export	5.000	34.727	contact fail
Fruits	existing export	0.0005	2,506.575	halted
Home coffee	existing export	5.000	17.755	contact fail
Kitchen	existing export	0.008	70.350	halted
Outdoor	existing export	5.000	25.988	contact fail
Toys	existing export	5.000	32.267	contact fail
Living room	selected derivative	5.000	1.825	contact pass

Table 11: Existing image-conditioned records for all nine public source-video families. The eight existing exports do not form one frozen benchmark, and the selected living-room derivative underwent scene-specific assistance. $P_{\max}$ for halted runs is the value reached at numerical failure, not a physically meaningful settled penetration. The 2 mm contact target is only one part of the joint gate. The companion audit pins source-video, conditioning-image, XML and metric hashes. No official SimFoundry method outputs were run or compared here.



Figure 9: Living-room image to selected native scene. Thirteen predicted targets include two hinges and two volumetric tetrahedral pillows. The exact corrected model completes five seconds of gravity with 1.825 mm peak enabled-contact penetration; mass and material values are priors.



Figure 10: Toy-tabletop image to selected rigid native scene. All 11 predicted targets have native representations; the corrected five-second gravity audit peaks at 1.690 mm. The uncorrected same-input export peaks at 35.830 mm.



Figure 11: Toy video to selected native scene. Multi-frame masks group ten foreground items and a tabletop; geometry and initial placement are mainly frame-zero conditioned. Its corrected five-second rollout peaks at 1.929 mm, versus 55.867 mm for the initial video-derived export.

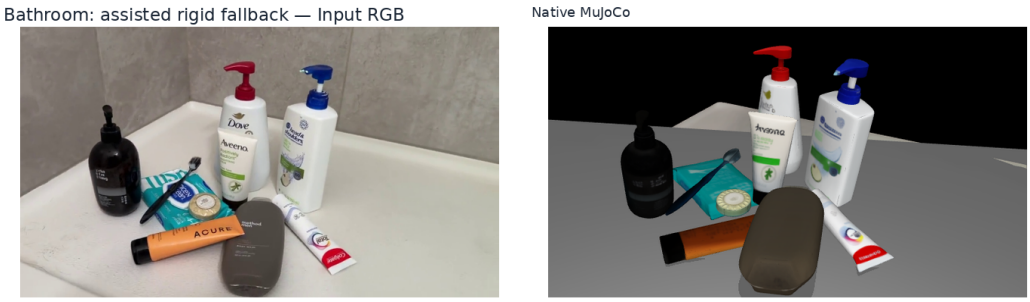

Figure 12: Bathroom image with an assisted *rigid fallback*. Four objects originally proposed as deformable are represented as rigid proxies, and same-scene contact parameters were tuned after inspection. The selected 11-target XML completes five seconds with 1.621 mm peak penetration. This panel does not demonstrate automatic soft reconstruction.

J CASE-LEVEL CORRECTION AND FAILURE DETAILS

Correction lineage. The displayed scenes were selected after inspection of development results. The living-room laptop split at local $z = -100$ mm, hinge axes, support and contact parameters were tuned against the same image and simulation gate; an untuned one-second laptop-contact probe had a 4.982 mm peak, whereas the selected five-second model has 1.825 mm. These runs differ in duration and are not a controlled efficacy estimate. For the toy video, the original five-second export had 55.867 mm penetration, whereas the derivative using an observed tabletop proxy and pairwise contact translation has 1.929 mm. Both native trajectories have zero warnings. The correction code is algorithmic, but variant selection was made during development and the full human trial/time budget was not recorded. A frozen new-input pipeline success rate therefore cannot be inferred from Table 3.

Additional development failures. A new dining-table image produces 11 native target representations, including a hinge candidate, but its initial state has a 34.727 mm collision penetration and fails the 2 mm gate despite a completed trajectory. A derivative contact-refinement attempt still exceeds 33 mm. The first-frame support geometry is also visibly incomplete. A second new image of a basket and nine plush toys looks recognizable at time zero and represents all 12 predicted targets, including nine contact-enabled cloth shells, but its initial export fails on the first 0.5 ms step; changing to bulk-soft proxies also fails on the first 2 ms step with many inverted tetrahedra. The shell geometry does not model the plush filling. These cases demonstrate why represented/predicted coverage and absence of solver warnings are insufficient to certify a usable scene. The interrupted plush runs are outside the 68 completed trajectories in Table 1. An earlier bread-rack video branch also failed object coverage and is retained as a diagnostic.

Two predeclared additional runs. The toaster-tabletop export has 7/7 predicted representations, including a tetrahedral bread group, but inverts one tetrahedron at 0.014 s with no numerical warning. The outdoor tabletop completes five seconds without warnings but has 25.988 mm peak penetration and only 9/10 predicted targets with type-consistent native representations: pliers have one hinge, the fabric sleeve is a surface cloth, and a predicted articulated tape measure remains a single rigid body because no split was fitted. The pliers' internal servo commissioning achieved only 0.448 of its inferred joint range. The compiled water-bottle body mass is 24.23 kg, illustrating the scale/material-prior failure even when an object is represented. Neither case is a physical-parameter recovery result; both are retained in the fixed two-input denominator.

Assisted rigid fallback diagnostic. On the failed bathroom input, we manually replaced four predicted deformable packets/tubes with rigid proxies while keeping the same reconstructed geometry, then selected whole-pair translation and stiffer contact parameters against that same scene. The unchanged cold run failed after 0.002 s. A rigid-only derivative completed five seconds but had 68.925 mm peak penetration; the successive pair-translation, floor-contact and default-contact derivatives had 5.250, 3.207 and 1.621 mm peaks. The final exact XML represents all 11 targets as one fixed fixture and ten free rigid bodies and completes 2,501 audited states with zero warnings. Same-input predicted-mask mean IoU was 0.537 for rigid-only and 0.495 after pair translation, unchanged by either contact-parameter edit. The contact metric improved while the visual alignment proxy worsened. This selected single-scene lineage demonstrates an executable corrected derivative, not an automatic deformable-scene success or an isolated estimate of general correction efficacy; its masks are not independent ground truth. Exact intermediate hashes and records are in `bathroom_contact_ablation.json`.