RFW-S: Everyday Deformable Objects with Independent Simulation Verification

Author information pending

Abstract

Everyday deformable objects require both broad semantic coverage and physical configurations that can be tested under interaction. We introduce **RobotFlyWheel-Soft (RFW-S)**, a dataset of 100,000 generated assets spanning 485 everyday categories across soft solids, cloth sheets, shells, garments, and bags. Construction connects reference appearances and geometry to structure-aware representations, executable simulator parameters, calibration feedback, and independent testing of the selected configuration. The intersection of physical verification and runtime checks retains 83.4% of the pool and all categories. In an expanded bag experiment, an independent stimulus rejects 3.87% of candidate evaluations that satisfy calibration, demonstrating the additional screening provided by testing beyond the fitted interaction. Garment comparisons characterize the effects of combined material and structural settings; a grasping application separates task completion from a deformation-derived integrity criterion. RFW-S associates semantic coverage with explicit configuration and behavioral evidence, supporting reproducible deformable-object studies within the documented solver and interaction scope.



Figure 1 Everyday deformable objects for robot interaction. RFW-S spans 485 categories across soft solids, cloth sheets, shells, garments, and bags. This conceptual rendering illustrates the resource's scope.

1 Introduction

A sponge compresses, a towel drapes, and a garment folds. Reproducing these interactions requires geometry, structural constraints, material response, and contact to work together. Robots encountering diverse household objects need reusable assets whose physical behavior can be executed and inspected.

Existing resources address different parts of this need. Large 3D collections provide visual breadth [1]; clothing datasets and deformable environments supply specialized geometry and interaction [2, 3]. ManiTwin validates rigid-object grasp proposals, PhysX-Omni generates assets across physical types, and Agentic Real2Sim reconstructs recorded episodes [4–6]. RFW-S organizes a broad everyday deformable taxonomy together with executable configurations and independent behavioral evidence for individual assets.

For a reusable simulation resource, these forms of evidence must remain associated with the asset that produced them. A visually plausible mesh can lack a suitable simulation representation, and a numerically stable rollout can still fall outside the intended response criteria. RFW-S therefore records both the configuration and the conditions under which it was tested. Its verification results describe behavior in the documented simulator; material parameters are effective simulation controls rather than measured properties of corresponding physical objects.

Three requirements guide its construction. *Semantic coverage* must remain visible after filtering: asset counts can conceal missing categories. *Structural compatibility* precedes parameter tuning: filled objects, open bags, and thin sheets require different representations. *Behavioral evidence* must describe the configuration actually executed, including its response beyond the interaction used for calibration.

RobotFlyWheel-Soft (RFW-S) contains 100,000 assets in 485 everyday categories and five structural families (Figure 1). Category-balanced generation establishes coverage; structure-aware construction selects a representation and effective solver controls. Calibration supports bounded corrections, after which a distinct probe tests the frozen candidate. Physical verification and a separate runtime check jointly retain 83,365 assets and every category. In controlled bag evaluations, independent testing rejects 325 calibration-accepted candidates across 8,400 evaluations, quantifying its additional screening.

Our contributions are:

- **A broad everyday deformable resource**: 485 categories across five structural families, with explicit pool and verified-subset statistics.
- **An integrated physical construction procedure**: generated geometry, structure-compatible representations, effective parameters, bounded calibration feedback, and independent final testing.
- **Quantitative usability evidence**: full-pool coverage and verification, garment response comparisons, parameter evaluations, correction records, and an application separating task completion from integrity.

2 Related Work

Resources and physical generation. Objaverse-XL provides broad visual coverage [1]. CLOTH3D, ClothesNet, and GarmentCodeData organize garment geometry [2, 7, 8]. SoftGym, PlasticineLab, DaXBench, and GarmentLab supply deformable interaction environments [3, 9–11]; BEHAVIOR-1K connects household objects and activities [12]. Table 1 compares resource units, coverage, physical construction, and evaluation evidence. ManiTwin-100K provides 512 rigid-object categories and simulation-verified grasps [4]. PhysXVerse spans over 2.9K general categories, supporting physical generation and behavior-based evaluation [5]. RFW-S focuses its taxonomy on everyday deformables and reports category retention after configuration-specific physical and runtime checks.

Physical inference and feedback. Language priors and video observations support parameter inference [13–15]. Vid2Sid and Swim2Real use visual feedback for system identification [16, 17]; Agentic Real2Sim assembles executable episodic twins [6]. RFW-S uses calibration feedback during asset construction, then records a separate final test. Its parameter comparisons use shared probes and evaluation budgets.

Table 1 Resources and their physical evidence. Reported units and category definitions differ across resources. Panel B describes the published construction and evaluation protocol. RFW-S tests a frozen configuration beyond its calibration interaction. Citations are given in the accompanying text.

A. Scale, semantic organization, and physical scope

Resource	Reported scale / unit	Category organization	Object scope
Objaverse-XL	10M+ 3D objects	Diverse source collections	Broad visual geometry
CLOTH3D	11.3K garments; 2.1M time samples	Six garment groups	Simulated clothing
SoftGym	10 environments	Task-defined objects	Rope, cloth, fluid
ManiTwin-100K	100K 3D objects	512 categories	Rigid graspable objects
PhysXVerse	8.7K+ 3D assets	2.9K+ general categories	Rigid, deformable, articulated
Agentic Real2Sim	100 rigid episodes evaluated; qualitative extensions	Episode-defined objects	Rigid, deformable, humanoid
RFW-S	**100K pool; 83,365 jointly verified**	**485 everyday categories**	**Five deformable families**

B. Parameters, validation protocol, and associated evidence

Resource	Physical construction	Reported evaluation	Evidence / reusable output
Objaverse-XL	Source 3D geometry	Visual-generation evaluation	Deduplicated objects, renders
CLOTH3D	Blender fabric presets	Cloth motion on body sequences	Meshes, vertex motion, fabric labels
SoftGym	Parameterized FleX environments	Manipulation-policy rollouts	Environments, state and reward
ManiTwin-100K	Inferred size, mass, friction	Grasp stability and slide resistance	Mesh, properties, verified grasp poses
PhysXVerse / Omni	Part-wise physical attributes	Human annotation review; free-fall / water-drop scoring	Physical assets, six-attribute benchmark
Agentic Real2Sim	Inferred episodic simulation	Replay-based candidate judging	Executable twins, replay observations
RFW-S	**Representations and effective solver values**	**Calibration, frozen-candidate independent probe, runtime check**	**Geometry, parameters, reports, correction traces**

Deformable manipulation and damage. DefGraspSim evaluates deformable grasps; stress-guided learning and OopsieVerse study fragile interaction and damage-aware outcomes [18–20]. TopoCut addresses damage-driven cutting [21]. Our grasping evaluation uses a deformation-derived damage proxy to expose the distinction between completing a task and preserving integrity.

Visual resources and simulator agents. Visual collections such as ABO, OmniObject3D, and Objaverse provide textured or annotated objects [22–24]. Simulation-oriented generation includes Seed3D, PhysX-Anything, and PhysX-Omni [5, 25, 26]; ManiTwin studies large-scale data-generation-ready objects [4]. Table 1 compares scale, coverage, and physical evidence. DeformableRavens supplies goal-conditioned tasks across cables, fabrics, and bags [27], complementing RFW-S's asset-level scope.

Simulator agents have generated tasks, scenes, rewards, and domain randomization [28–31]. RFW-S uses simulator observations to construct reusable physical assets and records corrections as well as non-acceptance. Its evidence concerns representation choice, executed parameter updates, independent verification, and category retention across the full pool.

3 The RFW-S Dataset

3.1 Coverage and resource units

The pool contains 100,000 assets in 485 canonical everyday categories. A category denotes an object type; a structural family groups representation requirements. The five families contain 168 soft-solid, 89 cloth-sheet,

63 shell, 94 garment, and 71 bag categories (Table 2). Appearance and parameter variations retain the same category identity. Generation assigns 206 assets to 395 categories and 207 to the remaining 90, providing near-uniform research coverage. After both checks, every category retains between 3 and 207 assets.

Four units recur in the analysis. A *category* is a semantic type, an *asset* is a generated object with a persistent identity, a *configuration* specifies its geometry, representation, parameters, and solver settings, and an *episode* executes a configured interaction. Repeated episodes and parameter variants can share the same underlying asset. Category counts describe semantic breadth; asset counts describe the pool; episode counts describe the amount of evaluation. We preserve these distinctions when reporting repeated method and budget comparisons.

Table 2 Full-pool census. Physical and runtime checks use the same 100,000 assets. "Both" is their configuration-matched intersection; all categories remain represented.

Family	Types	Pool	Physical	Runtime	Both
Soft solids	168	34,641	21,112	34,502	21,103
Cloth sheets	89	18,347	18,308	18,097	18,064
Shells	63	12,990	11,415	12,990	11,415
Garments	94	19,386	19,055	19,386	19,055
Bags	71	14,636	13,728	14,611	13,728
Total	**485**	**100,000**	**83,618**	**99,586**	**83,365**

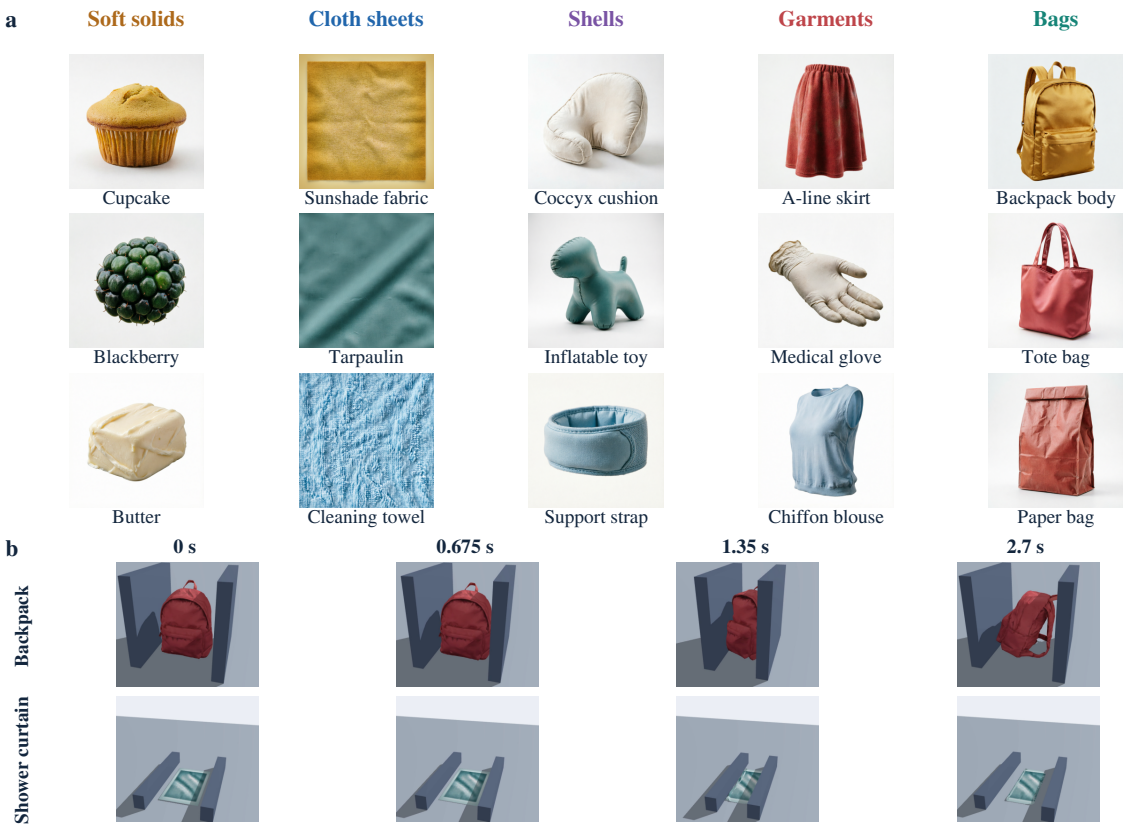


Figure 2 Reference appearances and recorded simulation. Fifteen reference appearances sample the unfiltered asset pool across its five structural families. Two recorded simulation sequences show compression and release at matched physical times within each sequence. Reference images convey intended appearance, while simulation frames show the executed representation.

3.2 Asset contents and provenance

Each asset links semantic identity and reference appearance to visual geometry, a simulation representation, scale, effective parameters, and verification records. Figure 3 shows a concrete example. Visual and physical meshes may differ in topology; their association and coordinate transforms are recorded. Numerical scale normalization is stored separately from intended semantic size. The physical record identifies the executed solver controls, including clamps and representation-dependent maps.

Reference appearances, generated visual meshes, and simulated proxies serve different purposes. Reference images specify intended appearance; the visual mesh provides the generated geometry; the physical representation supports the executed solver. Figure 2 shows reference images and recorded interactions, while Figure 3 links these stages for a single asset. The conceptual teaser summarizes the resource's scope and is not a recorded robot experiment.

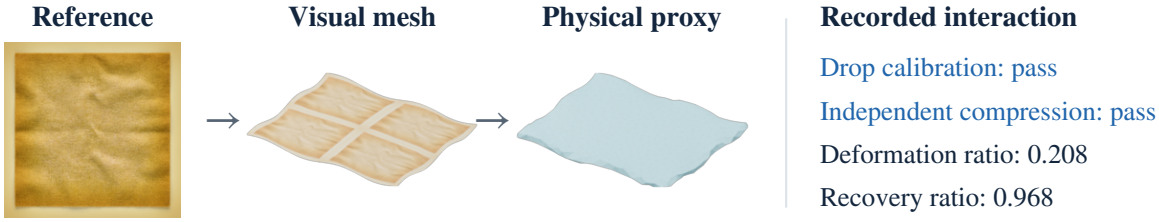


Figure 3 One asset from appearance to evidence. The reference, visual mesh, and physical-proxy render belong to the same sunshade-fabric asset. The last panel reports that asset's calibration and independent compression outcomes.

3.3 Verification scope and retained subsets

Physical verification evaluates calibration and independent behavioral criteria for the selected configuration. A separate *runtime check* evaluates execution and numerical behavior under its declared probe. The verified intersection matches asset and configuration identities:

$$\mathcal{D}_{\cap} = \{a \in \mathcal{D}_{\text{pool}} : V_{\text{physical}}(a) = 1 \ \wedge \ V_{\text{runtime}}(a) = 1\}. \tag{1}$$

The implementation uses Genesis 1.3.3 [32], with position-based dynamics (PBD) and material point method (MPM) representations. Verification denotes usability under these solver and interaction conditions. The parameters are simulator-effective values, inferred during construction.

The physical and runtime checks are separately evaluated sets. Passing one does not imply passing the other: 253 configurations pass only physical verification and 16,221 pass only the runtime check. Their intersection is a computed subset of the 100,000-asset pool, matched by asset and physical configuration identity. The full-pool accounting in Table 2 includes both retained and rejected configurations, making the loss of within-category coverage visible even when all category names remain present.

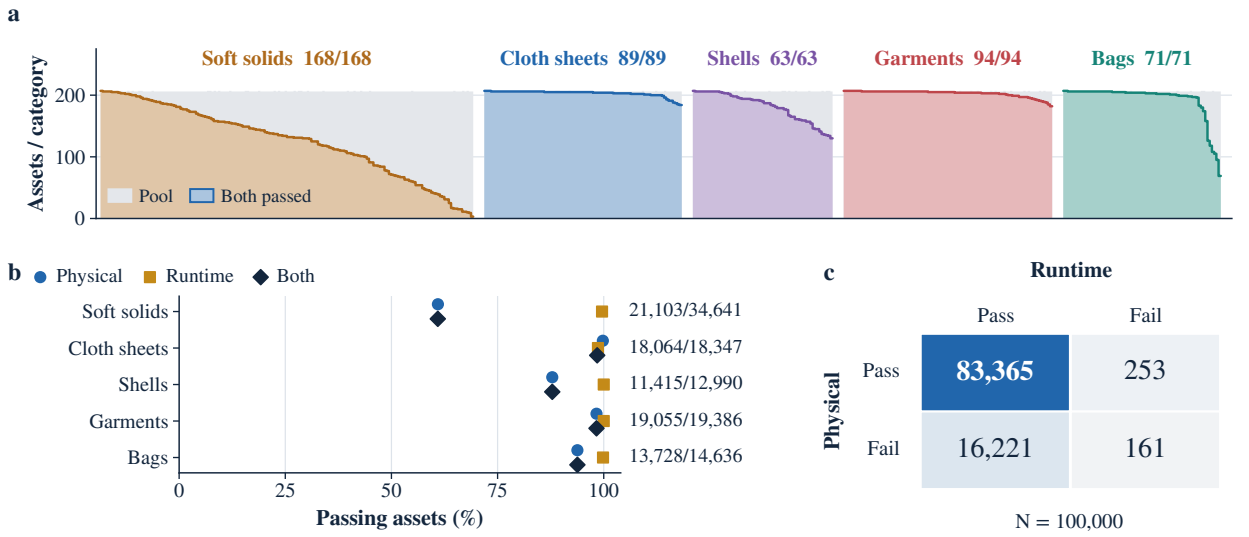


Figure 4 Coverage retained after verification. (a) Counts across all 485 categories. (b) Family pass proportions. (c) Exact overlap: 83,365 pass both checks, 253 only physical verification, 16,221 only runtime checking, and 161 neither. Counts are a full census.

4 From Generation to Verified Simulation

Construction separates representation choice, effective parameterization, candidate selection, and verification. Calibration observations can influence the candidate that is selected; the independent probe evaluates that selection after its geometry, parameters, and numerical settings are frozen. The separate runtime check is joined by configuration identity when forming the usable subset. This separation determines which evidence can support each acceptance decision.

4.1 Coverage-aware generation and structural adaptation

A canonical category and appearance specification condition Qwen-Image-2.0 reference generation [33]. References guide TRELLIS.2 geometry generation [34] and family-specific construction. Thin sheets can be procedural; structured objects retain their generated three-dimensional form. Geometry retries remain assigned to the same category and selected reference, maintaining the intended allocation before filtering (Figure 5).

Structure determines the representation: filled objects require interior material, bags need meaningful openings, and garments need surfaces that can fold. We use filled-volume, sheet, cloth-garment, shell-garment, bag-shell, reinforced-shell, and sealed-membrane profiles. Profiles determine compatible solvers and constraints; a filled beanbag can use MPM despite its semantic "bag" label. Volume construction uses interior filling and surface reconstruction. Surface adaptation removes small components and duplicated vertices while preserving openings. Connectivity, finite vertices, boundary edges, and repairs accompany the asset.

Semantic family guides this choice but does not uniquely determine a solver. An open fabric bag and a filled beanbag can require different representations despite sharing an everyday category family. The record therefore retains the proposed route and the mode actually executed. Representation diagnosis precedes interpretation of a parameter change: a control can only affect behavior when the selected solver and adapter implement it.

Garment adaptation addresses a recurrent failure: strong rest-curvature support can preserve a generated upright pose even with softer material. The flexible cloth profile reduces this support while retaining stretch resistance and self-collision. The matched configuration study in Section 5.2 measures the resulting behavior.

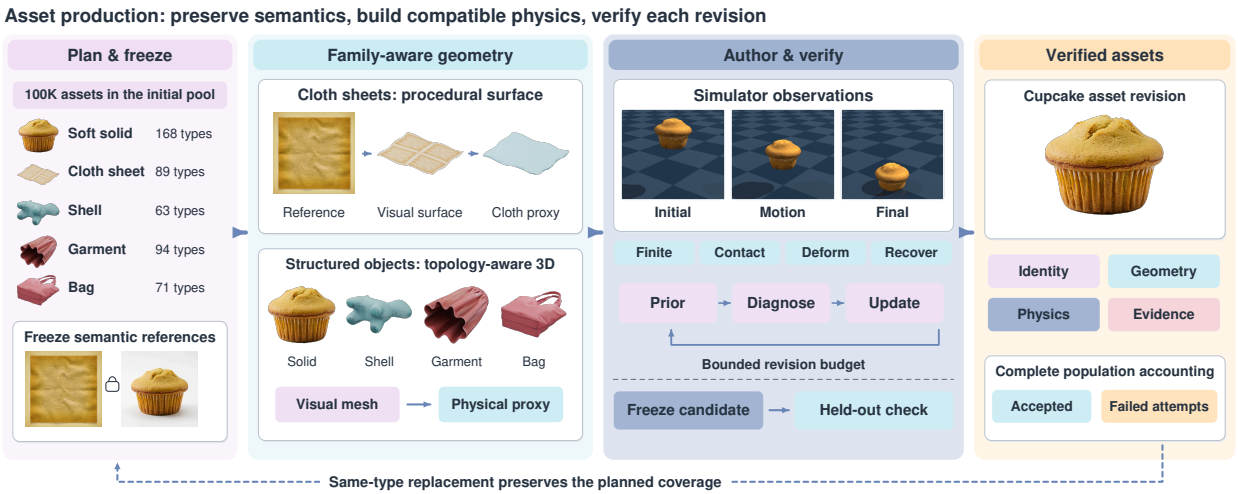


Figure 5 Resource construction and final verification. Coverage-aware generation supplies geometry and reference appearances. Structural adaptation and effective parameters make assets executable. Calibration supports corrections; independent testing and a separate runtime check provide evidence for the resulting configuration.

4.2 Effective parameters and calibration feedback

Given geometry x and representation r, Qwen (API identifier `qwen3.8-max`) estimates θ_0 from reference images and semantic cues. The adapter maps it to executable controls, $\theta_{\text{eff}} = A_r(\theta_0)$. A calibration probe $S_r(x, \theta_{\text{eff}}; u_{\text{cal}})$ measures contact, finite state, deformation, recovery, and residual motion where applicable (Figure 6). Diagnostics distinguish parameter mismatch, absent contact, numerical failure, and unsupported response. The same model proposes bounded corrections from structured calibration summaries, without image or video input to this correction stage. Iteration stops at acceptance or the candidate budget; Appendix F.1 gives the model settings.

Effective controls matter: a proposed viscosity field that is not passed to the MPM material leaves the response unchanged. Local diagnostics instead identify active controls, including yield stress for supported elastoplastic representations. Observation duration also matters: a short window can confound slow settling with persistent motion. The records therefore retain the executed parameter map, time step, substeps, and observation duration (Appendix B.2).

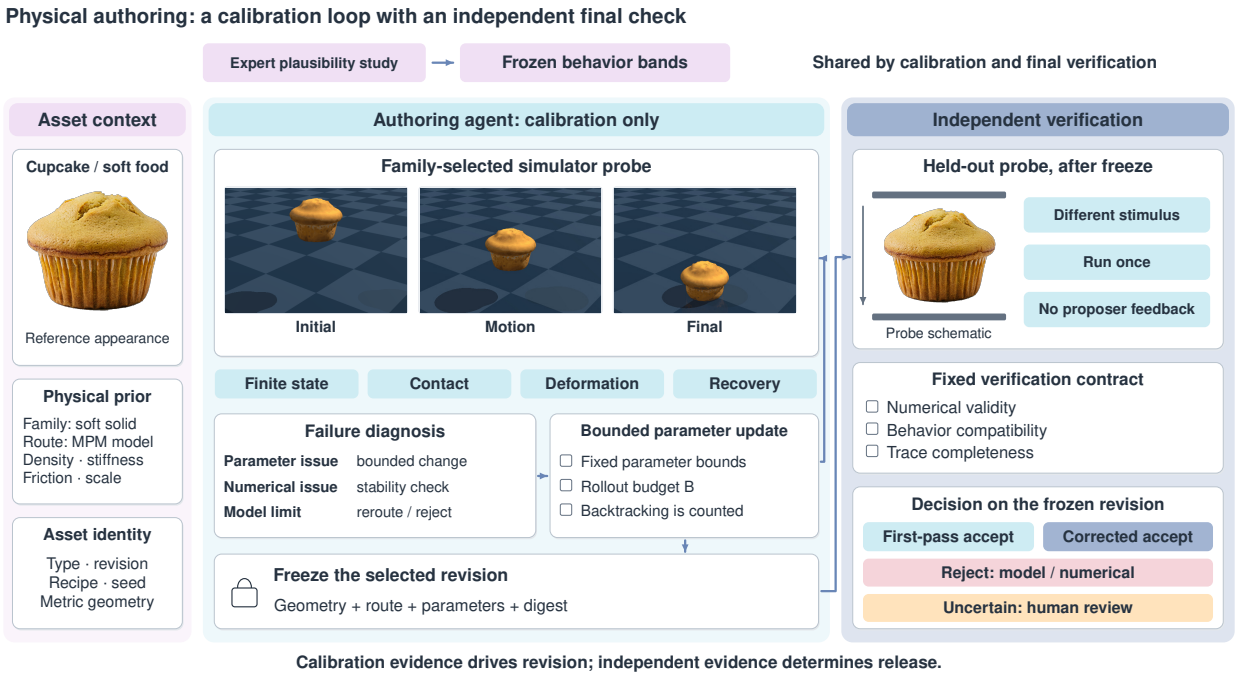


Figure 6 Calibration feedback and independent testing. Diagnostics guide bounded updates through supported solver controls. The selected geometry, route, and parameters are frozen before a distinct probe. Expert judgments inform behavioral criteria; numerical validity remains a separate requirement.

4.3 Model roles and configuration

Qwen-Image-2.0 generates reference appearances and TRELLIS.2 generates geometry. For physical construction, the recorded Qwen API identifier is `qwen3.8-max`. The physical-prior records store this same alias as the model revision; they do not identify an immutable checkpoint. The Qwen stages use API inference without task-specific fine-tuning. Table 3 separates image-conditioned initialization, text-only correction, and the numerical control. For the local text-only controls, this identifier is the configured default; the transport permits an override, and the canonical evaluation records do not store model identity for every request. These records therefore specify the available configuration without establishing per-request model provenance.

Table 3 Model roles in physical construction and local controls. Temperatures are sampling settings, not simulator parameters. The local controls' backend is specified by their transport configuration.

Component	Model / update	Input to proposer	Temp.
Physical prior	`qwen3.8-max`	Reference images, semantics, parameter catalog	0.1
Production correction	`qwen3.8-max`	Prior, current candidate, calibration metrics and diagnosis	0.2
Local text-only controls	`qwen3.8-max`	Structured parameter JSON; no images or video	0.2
Numeric agent	Mechanical rules	Numerical calibration residuals	—

4.4 Frozen-candidate verification and behavioral criteria

After calibration, geometry, representation, parameters, and numerical settings are fixed. A distinct probe evaluates the candidate, for example with lateral compression and release following drop calibration. Its result is not returned to the proposer for further updates to that candidate:

$$V_{\text{physical}} = G_{\text{execution}} \wedge G_{\text{cal}} \wedge G_{\text{independent}}. \tag{2}$$

Records distinguish first-candidate acceptance, acceptance after correction, model-limit rejection, numerical rejection, and unresolved judgment. A corrected acceptance requires an initial failure, a configuration change, and an independent pass.

Algorithm 1 summarizes the evidence boundary. An independent-test rejection remains an outcome of the frozen candidate; it is not fed back into another update within that candidate's verification. The separately scheduled runtime check contributes a second outcome, so the dataset intersection does not form a sequential filtering funnel.

Algorithm 1 Candidate selection and independent verification

Require: Asset geometry and context, calibration probe, independent probe, candidate budget
Construct a compatible representation and map the physical prior to effective controls
while calibration budget permits and the candidate is not accepted **do**
    Execute calibration; record the response, diagnosis, and complete configuration
    if calibration criteria are satisfied **then**
        Select the current candidate and stop calibration
    else if a supported correction and remaining budget are available **then**
        Apply a bounded update to supported physical controls; retain the update trace
    else
        Record the rejection or unresolved diagnosis and stop calibration
    end if
end while
if a calibration-accepted candidate was selected **then**
    Freeze its geometry, representation, parameters, and numerical settings
    Run the independent probe once; record the physical-verification outcome
end if
Associate the separately executed runtime outcome by configuration identity
return Asset record with distinct calibration, independent, and runtime evidence

Ten experts rated 150 core clips and 40 anchors, one rating per clip. Among 38 clips exceeding a residual-speed ratio of 0.05, 37 were judged settled. These observations motivate retaining settling as a diagnostic and using representation-dependent deformation and recovery bands. The ratings describe perceived plausibility; the study's protocol and per-dimension counts appear in Appendix C.

5 Experiments

We evaluate retained coverage, configuration-dependent response, the additional information from independent testing, and task versus integrity outcomes. Counts include failures; missing measurements remain explicit. Repeated method/budget runs share assets, and paired statistics use the asset as the resampling unit (Appendix F).

Table 4 lists the populations behind the results. The full-pool census and the smaller controlled studies answer different questions: coverage concerns generated assets, parameter comparisons concern repeated asset/target evaluations, and grasping concerns condition variants of a small set of base geometries.

Table 4 Evaluation populations and counting units. Repeated runs do not add new categories or base geometries. The two full-pool checks refer to the same asset pool and are reported separately.

Study	Distinct units and design	Evaluations
Full-pool verification	100,000 assets, 485 categories	100,000 per check
Supplementary stability	2,000 bags, stability drop	2,000
Garment configurations	4 garments × 3 settings × 2 probes	24
Response matching	30 asset/target pairs × 7 methods × 4 budgets	840
Expanded verification	300 bags × 7 methods × 4 budgets	8,400
Grasping	1,000 instances from 7 base geometries × 4 controllers	4,000
Expert judgments	150 core clips + 40 anchors; one rating per clip	190 ratings

5.1 Coverage survives verification, with uneven retention

Physical verification accepts 83.62% of the pool, runtime checking 99.59%, and their intersection 83.37% (Table 2). All 485 categories remain represented (Figure 4). The 16,221 assets that run successfully but miss physical criteria quantify the gap between execution and behavioral acceptance. Family retention ranges from 60.92% for soft solids to 98.46% for cloth sheets.

Soft solids account for 13,493 of 16,317 model-limit rejections (82.69%). Low-retention categories include gummy candy (3/206), nama chocolate (9/206), and toffee (10/207). These counts identify where the present representations leave sparse coverage. A targeted inspection of 200 rejected objects and 20 controls diagnoses representation failures; a separate 2,000-bag stability-drop check succeeds in 1,998 cases. Selection criteria and family outcomes are retained in Appendix A.

5.2 Garment configurations produce distinct responses

Four garments are tested under drop and compression with three configurations: *shape-preserving*, *softer material*, and *flexible cloth*. The 24 executions share source geometry and probe timing within each garment. Configurations jointly change material compliance and structural support. Table 5 gives both actions; Figure 7 shows the paired drop responses.

Table 5 Garment configuration comparison. Each cell averages four garments. Height is final/initial vertical extent; speed is the final maximum particle speed, averaged across runs.

	Drop		Compression	
Configuration	Height ratio	Speed (m/s)	Height ratio	Speed (m/s)
Shape-preserving	0.499	0.00091	0.876	0.15938
Softer material	0.470	0.03125	0.707	0.00571
Flexible cloth	0.312	0.14820	0.303	0.15665

The flexible configuration reduces mean drop height ratio from 0.499 to 0.312, versus 0.470 with softer material. Under compression the means are 0.876, 0.303, and 0.707, respectively. Its drop residual speed also rises from 0.00091 to 0.148 m/s. The combined settings permit greater collapse and longer-lasting residual motion. Individual values and exact settings appear in Appendix B.

This is a comparison of complete configurations on four objects. Because compliance, bending support, and some scene settings change together, the observed differences cannot be assigned to a single constraint. Height and residual speed characterize the recorded behavior; greater deformation alone does not establish closer agreement with a real garment.

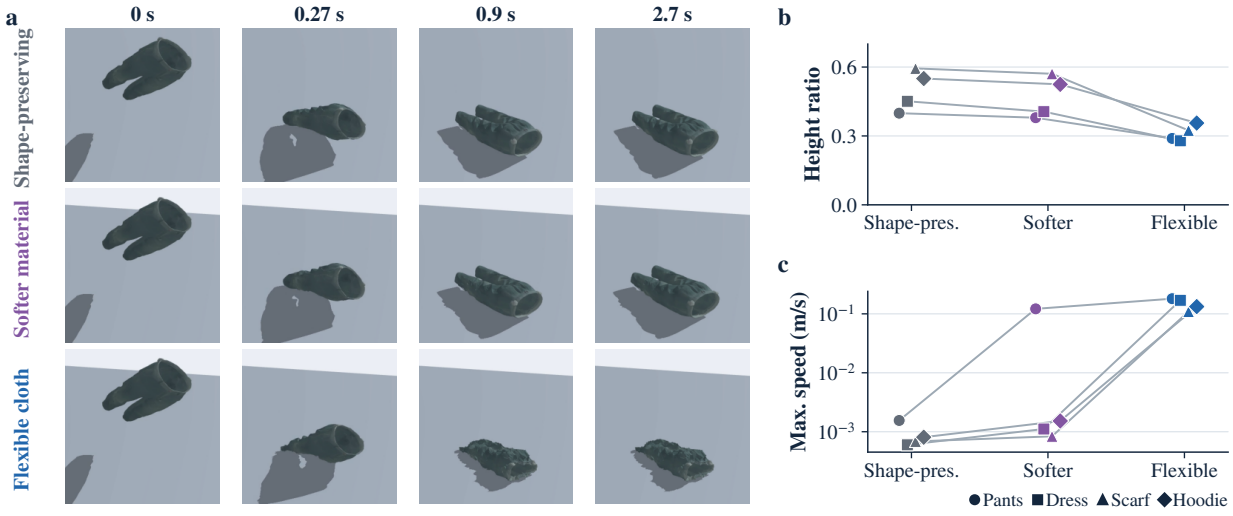


Figure 7 Deformation and residual motion under drop. (a) Matched physical times for one garment. (b,c) All four paired height and speed measurements; speed uses a log scale. Configurations match Table 5; Appendix B details the settings.

5.3 Independent testing reveals calibration-accepted mismatches

Evaluation protocol. The response-matching experiment uses 30 asset/target pairs, seven methods, and four allowed budgets $(1, 2, 4, 8)$, giving 840 runs. The expanded evaluation uses 300 bags with the same grid, giving 8,400 runs. Calibration compares the candidate's behavioral summary with the reference response using normalized, clipped differences; an independent stimulus evaluates the frozen candidate. The acceptance thresholds are 0.05 for calibration and 0.08 for independent response error (Appendix D). These response-matching criteria are distinct from the full-pool physical and runtime checks. The reference response is evaluated within the simulator. Consequently, the error measures agreement between simulated responses, rather than recovery of measured material constants.

For a candidate response c and reference response r, the adapter compares the available fields in

$$K \subseteq \{\text{finite fraction}, \text{contact fraction}, \text{settled}, \text{deformation ratio}, \text{recovery ratio}\}. \tag{3}$$

A Boolean mismatch contributes zero or one. A numerical field contributes a clipped relative difference. The recorded error is

$$L(c, r) = \frac{1}{|K|} \sum_{k \in K} d_k, \quad d_k = \begin{cases} \mathbf{1}[c_k \neq r_k], & \text{Boolean field,} \\ \min\left(1, \dfrac{|c_k - r_k|}{\max(|r_k|, 10^{-6})}\right), & \text{numerical field.} \end{cases} \tag{4}$$

Fields unavailable in either report are omitted. The adapter returns one if no fields can be compared. An infrastructure failure with no report remains missing in the exported error table. Success uses all scheduled runs as its denominator; mean error uses recorded values and is accompanied by their count.

Table 6 Calibration and independent-test outcomes. "Cal. only" passes calibration but fails the independent test; "Indep. only" has the reverse outcome. Counts are candidate evaluations across methods and budgets, with missing measurements kept separate.

Population	Both pass	Cal. only	Indep. only	Neither	Missing	Total
30 targets	682	49	3	83	23	840
300 bags	8,075	325	0	0	0	8,400

Additional screening. All 8,400 bag evaluations satisfy calibration, but 325 (3.87%) fail the independent stimulus (Table 6), involving 26 distinct bags. The distinct interaction thus exposes response mismatches after calibration acceptance. In the 30-target experiment, 49 of 731 observed calibration-passing evaluations fail independently; 23 runs lack measurements. The cross-tabulation quantifies the additional screening at the candidate level. The 8,400-evaluation denominator includes repeated methods and budgets; the 26-bag count reports distinct affected assets. These are two views of the same experiment, not independent estimates of full-dataset failure rates.

Parameter strategies. Table 7 compares fixed profiles, one-shot priors, rule-based updates, Sobol search, two text-only controls inspired by Vid2Sid and Swim2Real [16, 17], and the full numerical agent, which uses rules, backtracking, and abstention rather than a language-model update. Fixed profiles have lower pooled observed recovery error than the full agent. Expanded bag pass counts are close across methods: 1,148–1,159 out of 1,200 each. At budget 8, the paired full-agent minus fixed-profile error difference is -0.0036, with exploratory 95% interval [-0.0071, -0.0004]. Every expanded run evaluates one candidate, so these results characterize candidate outcomes. Budgets are separate executions; actual counts and paired intervals are provided in Appendix D.3.

The 30-target study records 772 one-candidate runs, 15 two-candidate runs, 30 three-candidate runs, and 23 runs with missing counts after infrastructure failures. All 8,400 expanded runs accept the initial candidate at calibration. The expanded study therefore tests independent screening of selected candidates; it does not exercise iterative parameter correction. Differences between its method-labelled runs come from separate executions of their initial candidates. The recorded comparisons do not establish a general benefit from larger search budgets or a superior correction policy.

Table 7 Seven-method independent evaluation. Recovery uses 30 targets and 120 runs per method; the expanded set uses 300 bags and 1,200 runs per method. Errors are means over recorded values (lower is better). All expanded errors are observed; recovery valid counts expose missingness. The two language-model controls are local text-only adaptations.

Method	Recovery pass	Error	Valid n	Expanded pass	Error
Fixed profile	100/120	0.0521	115	1152/1200	0.0127
One-shot prior	99/120	0.0704	119	1148/1200	0.0128
Rule-based update	98/120	0.0626	117	1151/1200	0.0125
Sobol search	97/120	0.0670	117	1156/1200	0.0121
Vid2Sid-style (text)	98/120	0.0622	115	1151/1200	0.0128
Swim2Real-style (text)	96/120	0.0734	118	1159/1200	0.0118
Full numerical agent	97/120	0.0691	116	1158/1200	0.0117

Observed feedback corrections. Production records contain 83,292 first-candidate acceptances and 326 acceptances after correction, including 209 cloth sheets. A filled-bag case changes density from 25 to 50 $\mathrm{kg/m^3}$ following a no-contact diagnosis. The revised drop passes, followed by independent compression with deformation ratio 0.120 and recovery ratio 1.038 (Figure 8). This trace links a contact diagnosis to an executed parameter correction and subsequent verification. The 326 corrected acceptances are observed production events. Estimating a causal recovery advantage would additionally require a controlled population of initial failures with matched starting configurations and update budgets.

Filled-bag correction	Initial drop		Revised drop	Independent compression
	FAIL	→	PASS	PASS
Density (kg/m³)	25		50	50
Contact fraction	0.000		0.524	1.000
Deformation ratio	0.000		0.142	0.120

Figure 8 A recorded correction and its independent test. Observed drop diagnostics lead to a density update. The revised candidate is frozen before compression. Values are recorded response summaries.

5.4 Task completion and a deformation-derived integrity criterion

A complementary grasp–lift–hold evaluation contains 1,000 qualified instances drawn from seven base geometries. Fixed command, slip feedback, a local force prior, and a privileged reference each evaluate the same 660 calibration and 340 held-out instances at one seed. The reference selects among five completed closure trials using privileged damage information. Its additional trials are reported separately.

The split separates instances, with base geometries shared across the two partitions. Held-out results therefore concern the reserved instances and conditions, rather than unseen object geometries. The privileged reference has access to damage readouts and additional trials; its outcome is reported separately from the single-trial controllers.

The damage score $D \in [0, 1]$ is derived from contact-time deformation-gradient statistics and configured onset/failure bands. Integrity is $1 - D$, and the damage flag is $D \geq 0.25$. Intact success requires both task completion and $D < 0.25$. This is a postprocessed response criterion, not a constitutively coupled fracture measurement. Appendix E gives the formula, sampling interval, fallback rules, and thresholds.

Table 8 Final grasping outcomes by instance split. Task success is shown for calibration and held-out instances. Intact success and median damage refer to the 340 held-out instances per method. Intact success requires task success and damage below 0.25. The last row is a privileged reference with five closure trials.

Controller	Cal. task	Held-out task	Intact	Damage
Fixed command	595/660 (90.2%)	245/340 (72.1%)	0/340	0.654
Slip feedback	634/660 (96.1%)	245/340 (72.1%)	0/340	0.784
Force prior (local)	600/660 (90.9%)	200/340 (58.8%)	0/340	0.478
Privileged reference	360/660 (54.5%)	200/340 (58.8%)	0/340	0.493

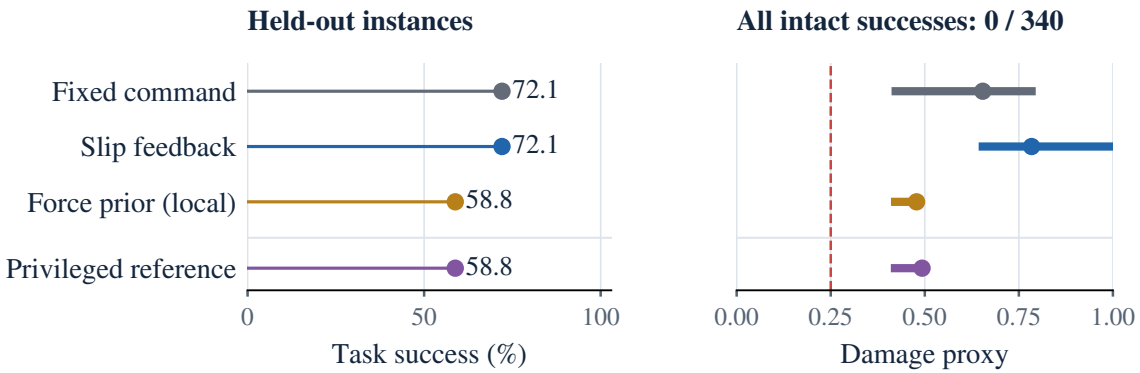


Figure 9 Task success and damage on held-out instances. Each controller evaluates 340 instances. Points give task success and median damage; horizontal segments show damage interquartile ranges. All instances exceed the damage threshold, so intact success is zero.

Across 4,000 recorded outcomes, 3,079 tasks succeed, but all 4,000 trigger the damage flag and none satisfies

intact success. Fixed command and slip feedback each complete 245/340 held-out tasks, despite different pooled success counts. Task completion and the specified integrity criterion thus give different assessments of the same outcomes (Table 8, Figure 9). Because every run crosses the damage threshold, binary intact success has no variation in this collection. The continuous damage distributions remain informative about the proxy response, but require calibration against external measurements before being interpreted as physical damage severity. Measured contact force is absent from the records; planned force, commanded force, and kinematic closure are therefore kept distinct.

6 Discussion and Limitations

6.1 Coverage, quality, and configuration-specific usability

RFW-S makes it possible to inspect semantic coverage together with the configuration and behavioral record of each evaluated asset. Retaining every category after verification preserves the breadth of the resource, while the unequal retained counts identify categories for which the current representations are less successful. Category retention is one aspect of dataset quality. Semantic correctness, within-category geometric diversity, and near-duplicate rates need their own audits. The targeted representation inspections reported here diagnose selected failures and do not estimate population-wide defect rates.

Verification is conditional on the executed solver, representation, parameters, and interaction. Changing any of these can change the response, so a previously accepted configuration should not be treated as a universal guarantee for other tasks. Likewise, simulator-effective parameters are not measurements of a corresponding real object's material constants. Paired physical-object measurements would provide an external reference for response fidelity and for the interpretation of inferred scale and material values.

6.2 Feedback evidence and independent verification

The independent probe adds information after calibration acceptance, and the cross-tabulation measures that contribution directly. Production traces also document successful corrections. These observations answer different questions: independent screening evaluates a selected candidate, whereas a controlled recovery study would compare update policies on the same initial failures. The present parameter comparisons are dominated by first-candidate executions. A stronger test of correction efficacy would use matched initial failures, trace prefixes for increasing budgets, and independent outcomes for each selected candidate. It should retain unsuccessful and abstained runs when computing recovery rates and evaluation cost.

The expert study informs plausibility criteria but supplies only one rating per clip. It therefore supports rating distributions and diagnostic examples, not inter-rater agreement estimates. Multiple ratings and a separate threshold-validation set would make the relationship between automated criteria and perceived plausibility more measurable. The archived case matrices use the settling-based decisions attached to their recordings; those labels are distinct from the final criteria, which retain settling as a diagnostic.

6.3 Interpreting the grasping application

The grasping study demonstrates execution and the joint reporting of task and integrity outcomes. Its damage score is a deformation-derived readout, and all evaluated instances exceed the chosen threshold. This saturation limits comparisons based on binary intact success. Calibrating the readout against physical measurements, testing a broader range of object geometries, and repeating conditions across seeds would strengthen its interpretation. The current instance split does not test generalization to unseen base geometries, and the privileged reference's extra trials must remain visible when comparing evaluation costs.

The resource supports further interaction and learning studies through reusable geometry, explicit parameter maps, and traceable verification records. The experiments here establish these construction and evaluation interfaces; they do not evaluate downstream policy-training gains. Future studies can use the same record structure to extend the stimulus set, add measured response references, and report how verification changes under new tasks and solver settings.

7 Conclusion

RFW-S connects 100,000 generated assets in 485 everyday deformable categories to structure-compatible representations, effective parameters, and recorded behavioral evidence. Physical verification and separate runtime checks jointly retain 83,365 assets and all categories. Independent testing exposes calibration-accepted response mismatches, while configuration comparisons and a grasping application distinguish deformation, task completion, and a specified integrity criterion. The resulting resource supports reproducible deformable-object studies with explicit configuration and evaluation scope. Broader stimuli and paired physical measurements can extend that evidence toward new interactions and real-object fidelity.

Reproducibility statement

The asset and physical manifests identify the geometry, representation, parameter configuration, and verification records used in each analysis. Tables are generated from the same per-run records used for the figures; failures and missing measurements remain in their denominators. Appendix F gives the analysis units and reproducibility procedure. The source package includes the derived tables, analysis scripts, and a manifest of figure inputs.

Ethics statement

The resource concerns generated models of inanimate objects. Its semantic coverage and visual priors can reflect biases in the source taxonomy and generative models. Expert judgments concern simulated behavior and are reported in aggregate. Simulator parameters and plausibility scores should be interpreted within the documented evaluation scope, particularly when considering transfer to physical robots.

AI use statement

Generative models are used for reference images, geometry, and parameter proposals. AI tools also assisted software development, literature lookup, and manuscript preparation. Quantitative results derive from recorded experiments. Reference appearances, simulation frames, and explanatory schematics are identified by their roles in figure captions.

References

[1] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10m+ 3d objects. In *Advances in Neural Information Processing Systems*, 2023.

[2] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: Clothed 3d humans. In *European Conference on Computer Vision (ECCV)*, 2020.

[3] Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. SoftGym: Benchmarking deep reinforcement learning for deformable object manipulation. In *Conference on Robot Learning (CoRL)*, 2020.

[4] 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*, 2026.

[5] Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-Omni: Unified simulation-ready physical 3d generation for rigid, deformable, and articulated objects. *arXiv preprint arXiv:2605.21572*, 2026.

[6] Guanxiong Chen, Qianjun Xia, Jiawei Peng, Heng Zhang, Pengyu Jing, Bole Ma, Justin Qian, Yixian Cheng, Ziyi Jiao, Bingyang Zhou, Yiduo Qu, Luoxin Ye, Kaifeng Zhang, Kunyi Wang, Weijia Zeng, Yunuo Chen, Pengzhi Yang, Ziqiu Zeng, Siyuan Luo, Huamin Wang, Chao Liu, Alan Yuille, Fan Shi, Changxi Zheng, Yunzhu Li, Chenfanfu Jiang, and Peter Yichen Chen. Agentic Real2Sim: Physics-based World Modeling with Vision-Language Agents. *arXiv preprint arXiv:2607.19190*, 2026. URL https://arxiv.org/abs/2607.19190.

[7] Bingyang Zhou, Haoyu Zhou, Tianhai Liang, Qiaojun Yu, Siheng Zhao, Yuwei Zeng, Jun Lv, Siyuan Luo, Qiancai Wang, Xinyuan Yu, Haonan Chen, Cewu Lu, and Lin Shao. ClothesNet: An information-rich 3d garment model repository with simulated clothes environment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20428–20438, 2023.

[8] Maria Korosteleva, Timur Levent Kesdogan, Fabian Kemper, Stephan Wenninger, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine-Hornung. GarmentCodeData: A dataset of 3d made-to-measure garments with sewing patterns. In *European Conference on Computer Vision (ECCV)*, 2024.

[9] Zhiao Huang, Yuanming Hu, Tao Du, Siyuan Zhou, Hao Su, Joshua B. Tenenbaum, and Chuang Gan. PlasticineLab: A soft-body manipulation benchmark with differentiable physics. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] Siwei Chen, Yiqing Xu, Cunjun Yu, Linfeng Li, Xiao Ma, Zhongwen Xu, and David Hsu. DaXBench: Benchmarking deformable object manipulation with differentiable physics. In *International Conference on Learning Representations (ICLR)*, 2023.

[11] Haoran Lu, Ruihai Wu, Yitong Li, Sijie Li, Ziyu Zhu, Chuanruo Ning, Yan Shen, Longzan Luo, Yuanpei Chen, and Hao Dong. GarmentLab: A unified simulation and benchmark for garment manipulation. *Advances in Neural Information Processing Systems*, 2024.

[12] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martin-Martin, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.

[13] Albert J. Zhai, Yuan Shen, Emily Y. Chen, Gloria X. Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang. Physical property understanding from language-embedded feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28296–28305, 2024.

[14] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. PAC-NeRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2303.05512.

[15] Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. PhysTwin: Physics-informed reconstruction and simulation of deformable objects from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7219–7230, 2025.

[16] Kevin Qiu, Yu Zhang, Marek Cygan, and Josie Hughes. Vid2Sid: Videos Can Help Close the Sim2Real Gap. *arXiv preprint arXiv:2602.19359*, 2026. URL https://arxiv.org/abs/2602.19359.

[17] Kevin Qiu, Kyle Walker, Mike Y. Michelis, Marek Cygan, and Josie Hughes. Swim2Real: VLM-Guided System Identification for Sim-to-Real Transfer. *arXiv preprint arXiv:2603.20827*, 2026. URL https://arxiv.org/abs/2603.20827.

[18] Isabella Huang, Yashraj Narang, Clemens Eppner, Balakumar Sundaralingam, Miles Macklin, Ruzena Bajcsy, Tucker Hermans, and Dieter Fox. DefGraspSim: Physics-based simulation of grasp outcomes for 3D deformable objects. *arXiv preprint arXiv:2203.11274*, 2022. URL https://arxiv.org/abs/2203.11274.

[19] Kei Ikemura, Yifei Dong, David Blanco-Mulero, Alberta Longhini, Li Chen, and Florian T. Pokorny. Sim-to-Real Gentle Manipulation of Deformable and Fragile Objects with Stress-Guided Reinforcement Learning. *arXiv preprint arXiv:2510.25405*, 2025. URL https://arxiv.org/abs/2510.25405.

[20] Arnav Balaji, Arpit Bahety, Sriniket Ambatipudi, Daniel Lam, Junhong Xu, and Roberto Martín-Martín. OopsieVerse: A Safety Benchmark with Damage-Aware Simulation for Robot Manipulation. *arXiv preprint arXiv:2606.31993*, 2026. URL https://arxiv.org/abs/2606.31993.

[21] Liquan Wang, Jiangjie Bian, Eric Heiden, and Animesh Garg. TopoCut: Learning multi-step cutting with spectral rewards and discrete diffusion policies. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 1282–1310, 2025. URL https://proceedings.mlr.press/v305/wang25z.html.

[22] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset

and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21126–21136, 2022.

[23] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 803–814, 2023.

[24] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023.

[25] Jiashi Feng, Xiu Li, Jing Lin, Jiahang Liu, Gaohong Liu, Weiqiang Lou, Su Ma, Guang Shi, Qinlong Wang, et al. Seed3D 1.0: From images to high-fidelity simulation-ready 3d assets. *arXiv preprint arXiv:2510.19944*, 2025.

[26] Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-Anything: Simulation-ready physical 3d assets from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2511.13648.

[27] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4568–4575, 2021.

[28] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. GenSim: Generating robotic simulation tasks via large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[29] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. RoboGen: Towards unleashing infinite data for automated robot learning via generative simulation. In *International Conference on Machine Learning (ICML)*, 2024.

[30] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[31] Yecheng Jason Ma, William Liang, Hung-Ju Wang, Yuke Zhu, Linxi Fan, Osbert Bastani, and Dinesh Jayaraman. DrEureka: Language model guided sim-to-real transfer. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. doi: 10.15607/RSS.2024.XX.094.

[32] Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond. https://github.com/Genesis-Embodied-AI/genesis-world, 2024.

[33] Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, et al. Qwen-Image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

[34] Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025.

A Asset Organization and Coverage

The dataset distinguishes a semantic category, a generated asset, a physical configuration, and an experimental episode. The census counts 100,000 completed geometric assets; parameter variants and episodes share their underlying asset identities. The intersection in Equation 1 is obtained by joining asset, physical, and runtime records and checking their configuration identities. The category distribution is supplied as a machine-readable table. Figure 10 illustrates the asset organization. The recorded examples in Figure 2 accompany the dataset description.

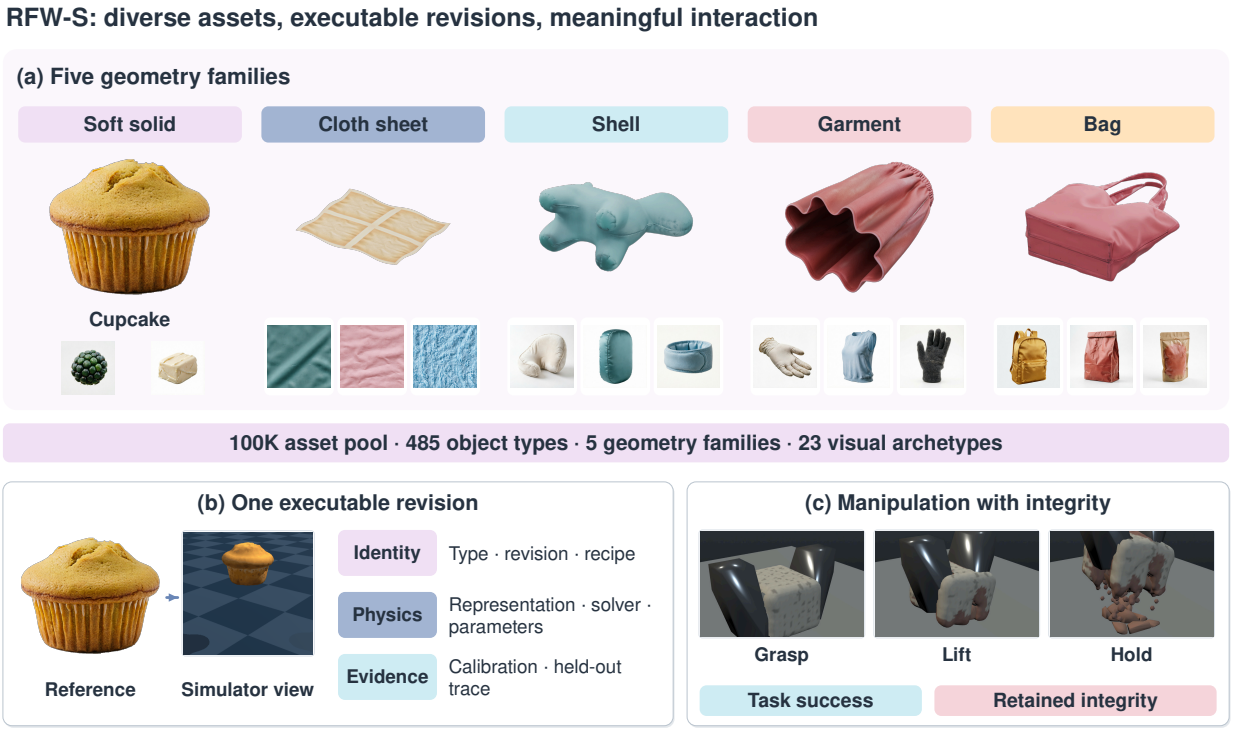


Figure 10 Asset diversity, executable configurations, and interaction. The five structural families connect semantic identity to geometry, physical parameters, and verification evidence. Reference appearances illustrate coverage, and the simulation schematics illustrate interaction and integrity criteria.

Table 9 Information associated with an evaluated asset configuration.

Component	Contents and purpose
Semantic identity	Canonical category, structural family, reference appearance
Geometry	Source mesh, visual appearance, simulation mesh, repair description
Scale	Intended size, solver normalization, coordinate transforms
Physical configuration	Representation, solver, proposed and effective parameters
Calibration	Probe configuration, measured response, diagnostic, candidate trace
Independent verification	Selected configuration, distinct probe, outcome and measurements
Runtime evidence	Separately executed check with matching configuration identity
Provenance	File fingerprints linking assets, configurations, and analysis inputs

The 300-asset routing analysis connects semantic identity to the solver mode reported by the executed calibration probe. It contains 193 PBD-cloth, seven elastic-MPM, and 100 elastoplastic-MPM executions. Table 12 reports their physical verification status. Each solver group contains the objects and representations assigned to it by the construction procedure. Figure 11 summarizes these checks and the expert ratings. The

Table 10 Construction outcomes across the complete asset pool. First-candidate acceptance and acceptance after correction are distinct events.

Construction outcome	Assets
Accepted at first candidate	83,292
Accepted after correction	326
Model-limit rejection	16,317
Numerical rejection	2
Unresolved judgment	63

Table 11 Construction outcomes by structural family. Corrected acceptances are events with a recorded initial failure and update; model-limit rejection is the recorded diagnosis.

Family	First pass	Corrected	Model limit	Numerical	Unresolved
Soft solids	21,071	41	13,493	0	36
Cloth sheets	18,099	209	36	0	3
Shells	11,349	66	1,553	1	21
Garments	19,055	0	330	1	0
Bags	13,718	10	905	0	3

initially proposed route and the executed mode are kept separately, since structural adaptation can change the route.

Table 12 Actual executed solver in the 300-asset routing sample.

Executed solver	Assets	Physically verified
PBD cloth	193	189
Elastic MPM	7	6
Elastoplastic MPM	100	74

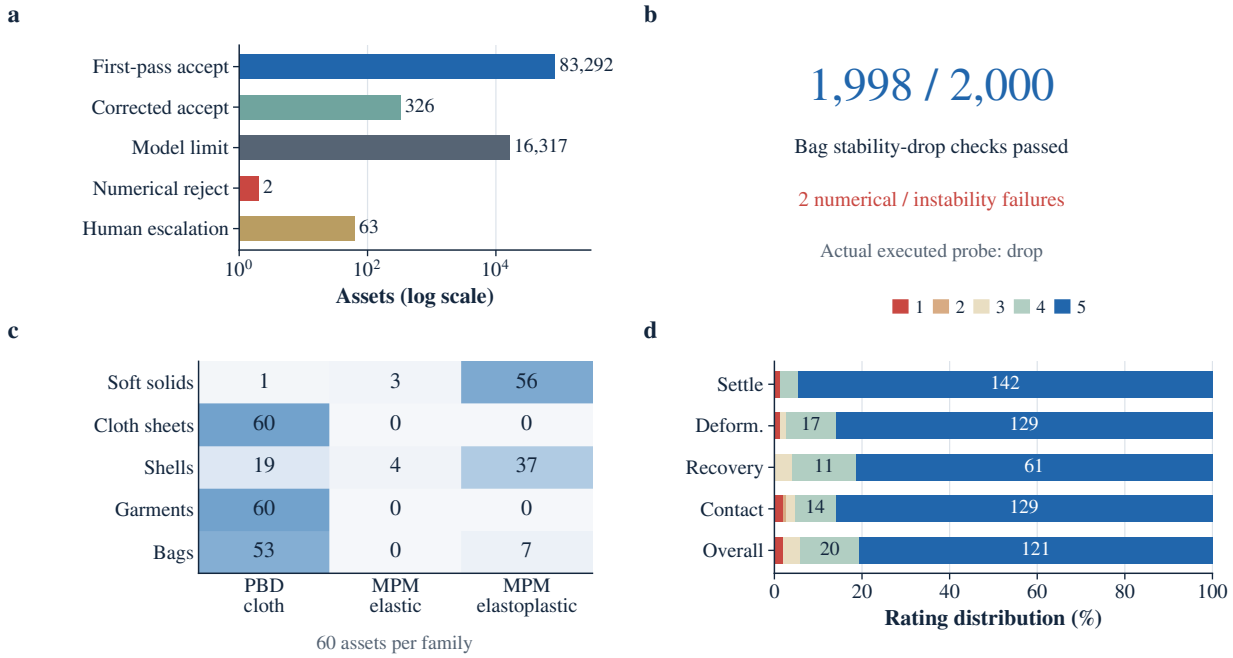


Figure 11 Supplementary verification evidence. (a) Production outcomes over 100,000 assets. (b) The 2,000 supplementary bag runs execute stability-drop checks, with 1,998 passes. (c) Executed solver modes for 300 assets, 60 per structural family. (d) Ratings of 150 core clips, with one expert rating per clip; recovery has 75 applicable ratings. Correction counts describe observed construction events.

B Structure and Effective Solver Controls

B.1 Matched garment configurations

The four garment cases are 000193, 000202, 002117, and 004801 in the asset index. Each is evaluated under drop and compression. The three configurations share the source geometry, initial pose, camera, and a 0.003 s step with ten substeps. Their material compliance, rest-curvature support, and floor extents differ. The three configurations are shape-preserving, softer material, and flexible cloth; their exact settings accompany the figure-source manifest.

The height ratio is final vertical extent divided by initial vertical extent. Final maximum speed is the largest recorded particle speed at the final sample, in m/s. The means in Table 5 summarize four objects each; Table 13 retains all paired measurements. Figure 12 visualizes the paired values. In particular, the mean speed can be dominated by one moving object, making individual values useful alongside the mean.

Table 13 All garment measurements. Each cell gives height ratio / final maximum speed (m/s).

Garment	Probe	Shape-preserving	Softer material	Flexible cloth
000193	Drop	0.399 / 0.002	0.379 / 0.122	0.289 / 0.181
000193	Compression	0.966 / 0.620	0.323 / 0.005	0.338 / 0.160
000202	Drop	0.451 / 0.001	0.406 / 0.001	0.279 / 0.170
000202	Compression	0.988 / 0.002	0.985 / 0.002	0.256 / 0.161
002117	Drop	0.594 / 0.001	0.571 / 0.001	0.324 / 0.110
002117	Compression	0.549 / 0.016	0.524 / 0.014	0.259 / 0.087
004801	Drop	0.550 / 0.001	0.525 / 0.002	0.356 / 0.133
004801	Compression	0.999 / 0.001	0.994 / 0.001	0.360 / 0.219

The broader representation inspection covers 220 targeted objects: 200 previously rejected examples and 20 controls. Source, drop, and compression views support inspection of each representation. This targeted

sample supports representation diagnosis. A separate geometric inspection identifies eight of 20 garments as suitable for local surface repair.

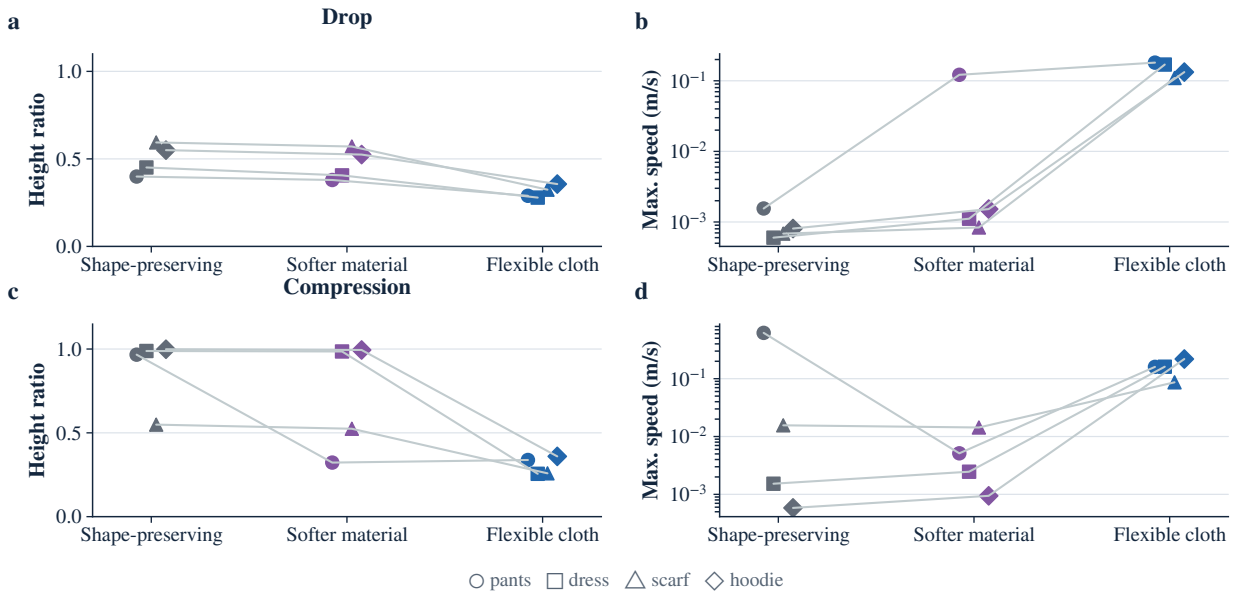


Figure 12 All 24 executed garment comparisons. Four garments are evaluated under three configurations and two actions. Lines pair the same garment across configurations. Height ratio and final maximum particle speed are separate response measures. The configurations jointly change compliance, bending iterations, and rest-curvature support.

B.2 Local parameter and observation-window studies

A local diagnostic study tests three soft solids and a cloth sheet (Figure 13). The adapter does not pass the proposed plastic-viscosity field to the MPM material. Altering that field therefore leaves the three solids' measured tail speeds effectively unchanged: approximately 0.1603, 0.0543, and 0.0853 m/s across the tested values. A supported yield-stress control changes the response; for example, the first solid's tail-to-peak speed ratio changes from 0.147 to 0.008 at the tested lower yield stress. The response is object-dependent: the second solid remains near the 0.05 diagnostic boundary. This motivates identifying effective solver controls before using a semantic parameter proposal as a correction direction.

Observation duration also changes the interpretation of residual motion. Extending the three solids' drop observations from 420 to 1,200 steps reduces their respective tail centroid speeds from 0.1603, 0.0543, and 0.0853 m/s to 0.0018, 0.0020, and 0.0121 m/s. The longer observation window resolves much of the residual motion in these three cases. The cloth checks likewise show sensitivity to the step/substep configuration. Production records retain executed step counts, time step, and numerical settings so that incomplete contact within a short observation window is distinguishable from an unsupported material response.

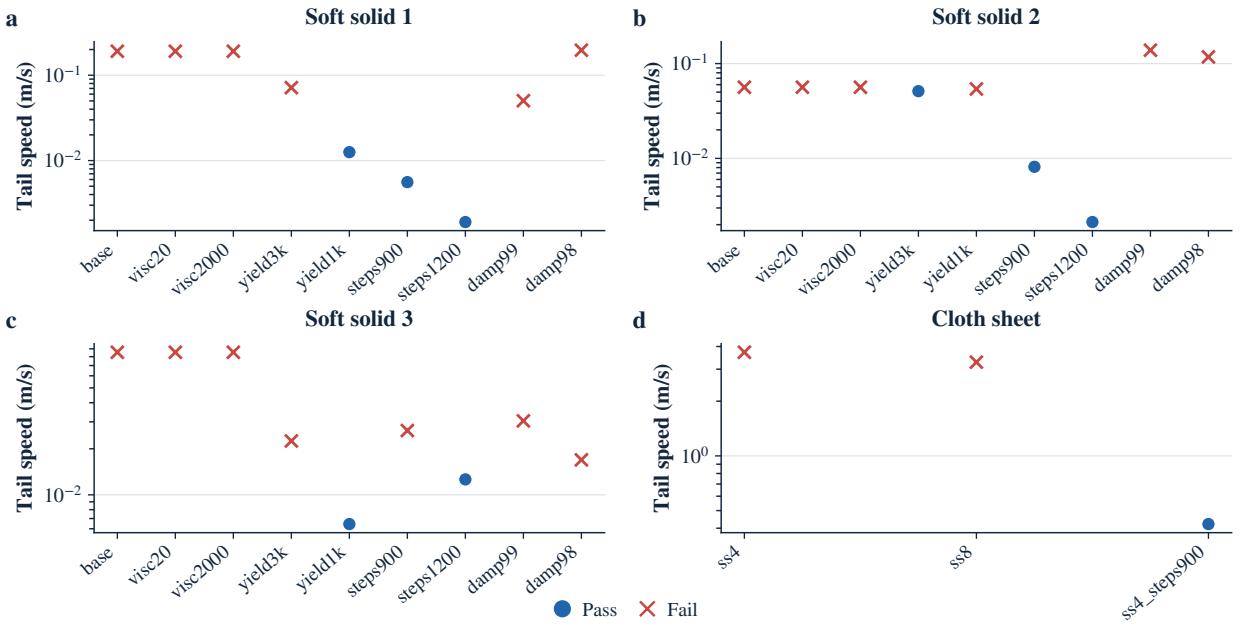


Figure 13 Local solver and observation diagnostics. Points show recorded tail mean particle speed. Colors and symbols indicate the complete behavioral decision, which also uses other measurements. Variants change material fields, damping, substeps, or observation duration; each point represents one case at its recorded settings, supplied with the data.

C Expert Plausibility Judgments

Ten experts each rated 19 clips, giving 190 annotations: 150 core clips and 40 anchors. Each clip received one rating. The core judgments cover settling, deformation, recovery where applicable, contact, and overall plausibility on five-point scales. Figure 14 illustrates the study and its use in defining behavioral criteria. Table 14 reports the number of ratings of at least four, with the applicable denominator for each dimension. Recovery has 75 applicable core clips; the other dimensions have 150.

Table 14 Core expert judgments. Each clip has one expert rating of perceived behavioral plausibility. Proportions use core clips only, with dimension-specific denominators.

Dimension	Rated clips	Score 4 or 5	Proportion
Settling	150	148	98.7%
Deformation	150	146	97.3%
Recovery	75	72	96.0%
Contact	150	143	95.3%
Overall	150	141	94.0%

The judgments identify useful distinctions between numerical diagnostics and perceived behavior. Among 38 clips exceeding a tail-to-peak speed ratio of 0.05, 37 were judged settled. Recovery could also remain plausible below a single elastic-style threshold. The adopted criteria therefore retain settling as a diagnostic and use representation-dependent deformation and recovery ranges. Finiteness, contact, and penetration checks are evaluated separately. The judgments guide plausibility criteria, with one rating per clip.

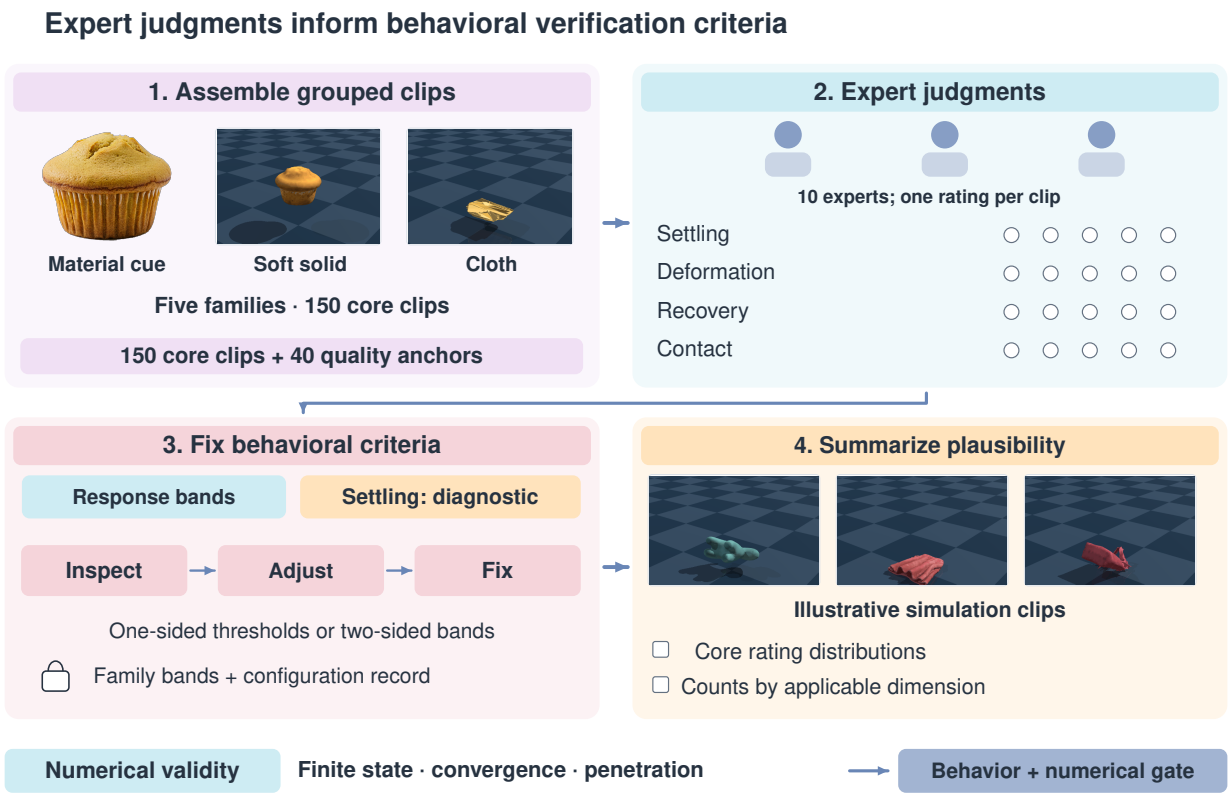


Figure 14 Expert plausibility judgments and behavioral criteria. Ten experts supplied one rating per clip for 150 core clips and 40 anchors. Judgments inform representation-dependent response bands; settling remains a diagnostic. Numerical validity is checked separately; the illustrated clips explain the assessment workflow.

D Parameter Evaluation and Statistics

D.1 Recorded error and acceptance

The response error $L(c, r)$ is defined in Equation 4. Success uses the scheduled-run denominator; mean error uses recorded measurements, with valid counts reported separately.

The recovery evaluation includes 30 assets and 30 target handles. Each method/budget pair schedules those 30 targets; the four budgets are separate runs. Of 840 runs, 772 evaluate one candidate, 15 evaluate two, 30 evaluate three, and 23 lack this measurement following infrastructure failures. Figures 15 and 16 show acceptance, error, and actual candidate counts alongside the allowed budgets. The expanded bag evaluation schedules 300 assets per method/budget pair, and all 8,400 runs evaluate one candidate. Its independent check accepts 8,075 candidates and rejects 325.

Table 15 Recovery error by allowed candidate budget. Parentheses give the number of recorded errors out of 30 scheduled targets.

Method	Budget 1	Budget 2	Budget 4	Budget 8
Fixed profile	0.0533 (28)	0.0527 (29)	0.0446 (29)	0.0579 (29)
One-shot prior	0.0698 (30)	0.0696 (30)	0.0836 (30)	0.0582 (29)
Rule-based update	0.0659 (29)	0.0727 (29)	0.0585 (29)	0.0537 (30)
Sobol search	0.0676 (29)	0.0583 (29)	0.0663 (29)	0.0757 (30)
Vid2Sid-style (text)	0.0756 (29)	0.0709 (29)	0.0438 (28)	0.0578 (29)
Swim2Real-style (text)	0.0633 (29)	0.0884 (30)	0.0658 (29)	0.0755 (30)
Full numerical agent	0.0576 (28)	0.0834 (29)	0.0734 (30)	0.0614 (29)

Table 16 Expanded bag verification error by allowed budget. Every cell contains 300 recorded values; all runs evaluate one candidate.

Method	Budget 1	Budget 2	Budget 4	Budget 8
Fixed profile	0.0126 (300)	0.0136 (300)	0.0104 (300)	0.0141 (300)
One-shot prior	0.0138 (300)	0.0126 (300)	0.0127 (300)	0.0119 (300)
Rule-based update	0.0135 (300)	0.0122 (300)	0.0129 (300)	0.0114 (300)
Sobol search	0.0128 (300)	0.0110 (300)	0.0120 (300)	0.0125 (300)
Vid2Sid-style (text)	0.0168 (300)	0.0109 (300)	0.0116 (300)	0.0119 (300)
Swim2Real-style (text)	0.0131 (300)	0.0115 (300)	0.0123 (300)	0.0104 (300)
Full numerical agent	0.0123 (300)	0.0121 (300)	0.0117 (300)	0.0105 (300)

Method labels describe the local implementation. In the plots, RuleTune is the rule-based update and Seeded Sobol is Sobol search. Vid2Sid-style and Swim2Real-style denote local text-only parameter controls; the latter adds backtracking. They use our parameter interface without the original methods' visual inputs. Numeric agent denotes rule-based updates with backtracking and abstention, with no language model in its update rule. Fixed profiles and one-shot priors can share initial values under this adapter. The numerical-agent comparison fixes the representation, while production also includes structural adaptation and construction decisions. Model configurations appear in Appendix F.1.

D.2 Calibration and independent-test cross-tabulation

The calibration response tolerance is 0.05 and the independent response tolerance is 0.08 in the executed evaluation adapter. A run with missing calibration or independent error is listed separately. The cross-tabulation joins both measurements within the same candidate evaluation, and the independent-pass column follows its recorded final outcome. In the 30-target experiment, 685 independent passes include three calibration failures; 682 satisfy both response criteria. The expanded experiment has 8,075 joint passes and 325 calibration-only passes involving 26 distinct bags.

Table 17 Cross-tabulation by method. "Both" passes both response criteria; the two "only" columns pass exactly one. Missing executions remain outside the four observed outcome cells.

Experiment	Method	Both	Cal. only	Indep. only	Neither	Missing
Recovery	Fixed profile	99	4	1	11	5
Recovery	One-shot prior	98	5	1	15	1
Recovery	Rule-based update	97	6	1	13	3
Recovery	Sobol search	97	9	0	11	3
Recovery	Vid2Sid-style (text)	98	10	0	7	5
Recovery	Swim2Real-style (text)	96	10	0	12	2
Recovery	Full numerical agent	97	5	0	14	4
Filter	Fixed profile	1152	48	0	0	0
Filter	One-shot prior	1148	52	0	0	0
Filter	Rule-based update	1151	49	0	0	0
Filter	Sobol search	1156	44	0	0	0
Filter	Vid2Sid-style (text)	1151	49	0	0	0
Filter	Swim2Real-style (text)	1159	41	0	0	0
Filter	Full numerical agent	1158	42	0	0	0

The methods share initial parameter values within each of the 120 recovery and 1,200 expanded asset/budget groups. Each allowed budget is executed as a separate run with its own candidate trace. Actual candidate counts accompany each method comparison.

D.3 Paired analysis and missingness

The analysis unit is an asset/target, with methods paired within each budget. For a paired error difference, both methods must have a recorded value. We draw 5,000 bootstrap resamples of asset IDs with replacement, using seed 20260926 and the same resampling indices across methods and budgets within each experiment.

Each replicate averages the available paired differences; its 2.5th and 97.5th percentiles define the reported interval. The count of observed pairs is reported explicitly. This retains the population's missing pair positions while conditioning the error estimate on observed pairs.

Table 18 Exploratory paired differences: full numerical agent minus fixed profile. Negative values favor the numerical agent on recorded error. Intervals are unadjusted for multiple comparisons.

Experiment	Budget	Pairs	Difference	95% interval
Recovery	1	27	-0.0102	[-0.0355, 0.0113]
Recovery	2	28	0.0227	[-0.0028, 0.0593]
Recovery	4	29	0.0196	[-0.0015, 0.0544]
Recovery	8	28	0.0052	[0.0000, 0.0117]
Filter	1	300	-0.0003	[-0.0043, 0.0037]
Filter	2	300	-0.0015	[-0.0048, 0.0018]
Filter	4	300	0.0013	[-0.0026, 0.0056]
Filter	8	300	-0.0036	[-0.0071, -0.0004]

Intervals and point estimates vary across budgets and experiments. The expanded experiment evaluates one candidate per run; the recovery experiment includes multi-candidate runs. Reporting actual candidate counts alongside success, error, and paired differences makes this distinction explicit.

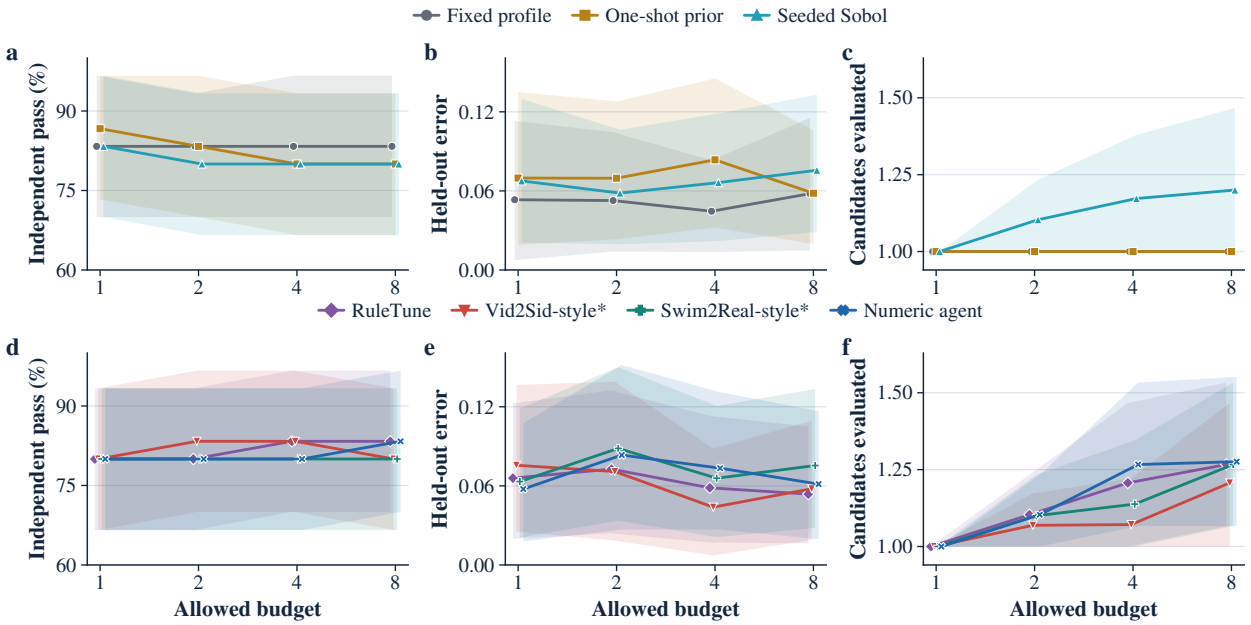


Figure 15 Recovery evaluation across allowed budgets. Columns show independent-pass fraction, observed behavior-summary error, and candidates evaluated; rows separate method groups. Budgets are separate runs. The 840 runs include 685 successes and 23 infrastructure failures with missing errors. Failures remain in success denominators. Whiskers are 95% percentile intervals from 5,000 asset/target bootstrap resamples with shared draws across methods and budgets. Asterisks mark local text-only language-model adaptations.

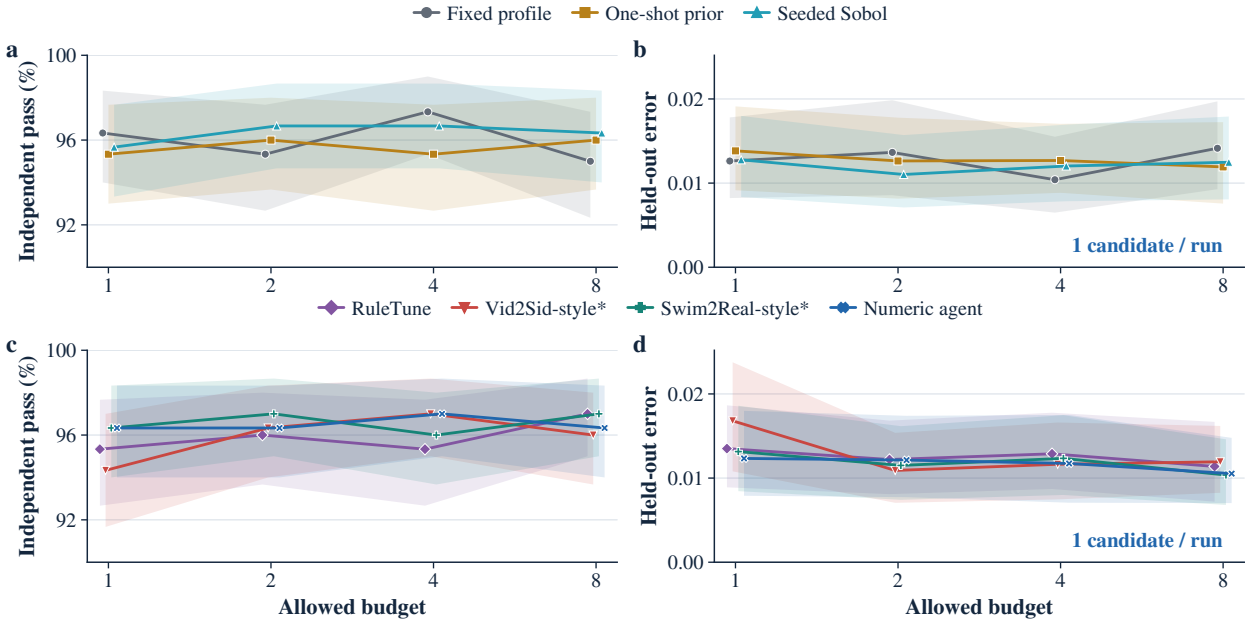


Figure 16 Expanded bag verification across allowed budgets. The 8,400 runs cover 300 bags, seven methods, and four allowed budgets; 8,075 pass independent verification. All errors are recorded and every run evaluates one candidate. Columns show pass fraction and response error; rows separate method groups. Bootstrap intervals follow Figure 15. Each point summarizes candidate outcomes at the indicated allowed budget.

E Complementary Grasp Evaluation

The construction produces 2,000 candidate instances, of which 1,000 qualify for evaluation. The instances comprise parameter and condition variants of seven base geometries. Each controller uses the same 660 calibration and 340 held-out instances at seed zero. The split partitions instances, including variants of shared base geometries.

Controllers and evaluation access. Fixed command follows the prescribed closure. Slip feedback tightens the closure in response to kinematic sag; the local force-prior controller maps its prior to the executable closure interface. Measured contact force is unavailable in all 4,000 records. Nominal force conditions, commanded force, and kinematic closure are therefore kept distinct. The privileged reference executes five complete closure trials with ratios 0.10, 0.20, 0.30, 0.45, and 0.60. It selects the least-damaged intact success when available, otherwise the least-damaged completed trial. The separate reference row reports the outcome of this five-trial selection using privileged damage readouts.

Deformation-derived damage readout. Let $F_i(t)$ denote the deformation gradient of particle i, and let $\mathcal{T}_c$ contain sampled observation times at which contact is detected. Every twentieth observation with contact contributes to

$$q = \max_{t \in \mathcal{T}_c} \mathrm{Q}_{0.95,i}\big(\|F_i(t) - I_3\|_F\big), \qquad D = \mathrm{clip}_{[0,1]}\left(\frac{q-a}{b-a}\right), \tag{5}$$

where I_3 is the identity matrix and (a, b) are the onset/failure bands. Low, medium, and high fragility settings use $(0.35, 0.75)$, $(0.25, 0.60)$, and $(0.15, 0.45)$, respectively. The implementation falls back to final mean deformation if no positive contact-time peak is available, and then to an aperture/extent statistic if the deformation gradient is unavailable. Thus D is a normalized deformation proxy derived from the simulated response.

Integrity is $1 - D$. A damage flag is triggered when $D \geq 0.25$; intact success requires task completion and

$D < 0.25$. Task success requires the recorded lift and hold criteria and a finite simulated state. The proxy is computed from the executed MPM/PBD response as an evaluation readout.

Table 19 All application splits. Task, intact success, and damage flag are counts; damage and integrity are medians. The flag overlaps task success. "All" pools the two instance splits; each controller has the same underlying instances.

Split	Controller	N	Task	Intact	Flag	Damage	Integrity
Calibration	Fixed command	660	595	0	660	0.658	0.342
Held-out	Fixed command	340	245	0	340	0.654	0.346
All	Fixed command	1000	840	0	1000	0.654	0.346
Calibration	Slip feedback	660	634	0	660	0.950	0.050
Held-out	Slip feedback	340	245	0	340	0.784	0.216
All	Slip feedback	1000	879	0	1000	0.784	0.216
Calibration	Force prior (local)	660	600	0	660	0.503	0.497
Held-out	Force prior (local)	340	200	0	340	0.478	0.522
All	Force prior (local)	1000	800	0	1000	0.489	0.511
Calibration	Privileged reference	660	360	0	660	0.501	0.499
Held-out	Privileged reference	340	200	0	340	0.493	0.507
All	Privileged reference	1000	560	0	1000	0.493	0.507

All 4,000 records contain valid damage and integrity measurements and report convergence. There are 3,079 task successes and 921 task failures; all 921 failures have the mutually exclusive damage-failure reason. Independently, the damage flag is present in all 4,000 outcomes, and intact success is zero. Failure reasons partition task failures, while damage flags apply to all outcomes. Across all instances, fixed command, slip feedback, force prior, and the privileged reference have 840, 879, 800, and 560 task successes, respectively. The held-out results are 245, 245, 200, and 200 out of 340. The split-specific results distinguish the equal held-out performance of the first two controllers from their different pooled counts. This complementary collection evaluates the construction interface and its task and integrity measures.

F Reproducibility and Implementation Details

F.1 Model roles and inference configuration

Section 4.3 and Table 3 identify the models, input modalities, and sampling temperatures. The following details specify the prompt and budget configuration.

Prompts and parameter constraints. The physical-prior configuration sets an 8,192-token output limit and disables the API's thinking option. Its prompt version is `softsim-physical-production-v1-qwen3.8-max`. Reference images accompany structured semantic and parameter-catalog information. Outputs are uncertain simulation priors, including parameter values, confidence, and rationale. Production correction uses a separate JSON-only prompt requesting one physical-parameter patch and a rationale from the current prior, candidate, and calibration metrics/diagnosis. The adapter restricts changes to supported physical parameters and enforces bounds; solver time step, substeps, and other numerical settings remain fixed. The source package records the prompt text, selected configuration fields, and their source fingerprints.

Budgets and local comparison scope. The production configuration starts with three candidates and allows extensions by two, up to eight, while relative calibration improvement meets the 0.05 threshold. Independent verification follows candidate selection. The parameter comparisons instead use separately scheduled budgets of 1, 2, 4, and 8; these bound simulator candidate evaluations, not API calls. The text-only controls share the configured Qwen transport, whereas the Numeric agent uses mechanical updates, backtracking with factor 0.5, and abstention after two consecutive non-improving updates. The released evaluation records count simulator candidates, so those counts should not be interpreted as language-model call counts. In particular, all 8,400 expanded runs accept the initial candidate at calibration and do not exercise a language-model correction. These results characterize candidate verification under the local implementations, not the original visual methods.

F.2 Records and analysis reproduction

Inputs and record identity. The census joins the asset manifest, physical-configuration index, and full runtime records. Asset and physical-configuration identities match across these inputs. The five complete evaluations contain 100,000 runtime records, 2,000 supplementary stability checks, 840 recovery records, 8,400 expanded verification records, and 4,000 application records. Completion metadata, task manifests, and canonical records have matching counts and recorded file fingerprints. Rejected and infrastructure-failure records remain present.

Analysis reproduction. The accompanying analysis package supplies machine-readable category and family counts, construction states, per-run parameter measurements, application outcomes, and their sources. The manuscript's table generator produces the result macros and tables from these files. Paired statistics use the same resampling scheme as the visualization scripts. Figure inputs record the source asset or episode, simulation configuration, file fingerprint, and frame time when applicable. The source package preserves explicit figure identities and includes the final figure files, numerical inputs, and their fingerprints.

Scope of auxiliary evidence. The 2,000-record supplementary check evaluates bag stability under drop. The structure study, expert judgments, local parameter diagnostics, and application collection each use their specified selection criteria and denominators, retaining infrastructure failures in the scheduled-run counts. Software fixtures and execution pilots are tracked separately.

G Qualitative Probe Outcomes

Figures 17 and 18 show eight completed drop and compression–release runs across the five structural families. PASS/FAIL reproduces the automated behavior decision attached to each recorded video. All four FAIL cases retain motion at the end of the observation window and fail its settling criterion. The labels refer to the settling-based assessment used for these recordings; the final acceptance criteria instead treat settling as a diagnostic (Appendix C). Each row shows a distinct asset and configuration, selected to illustrate passing or non-settling behavior.

G.1 Drop

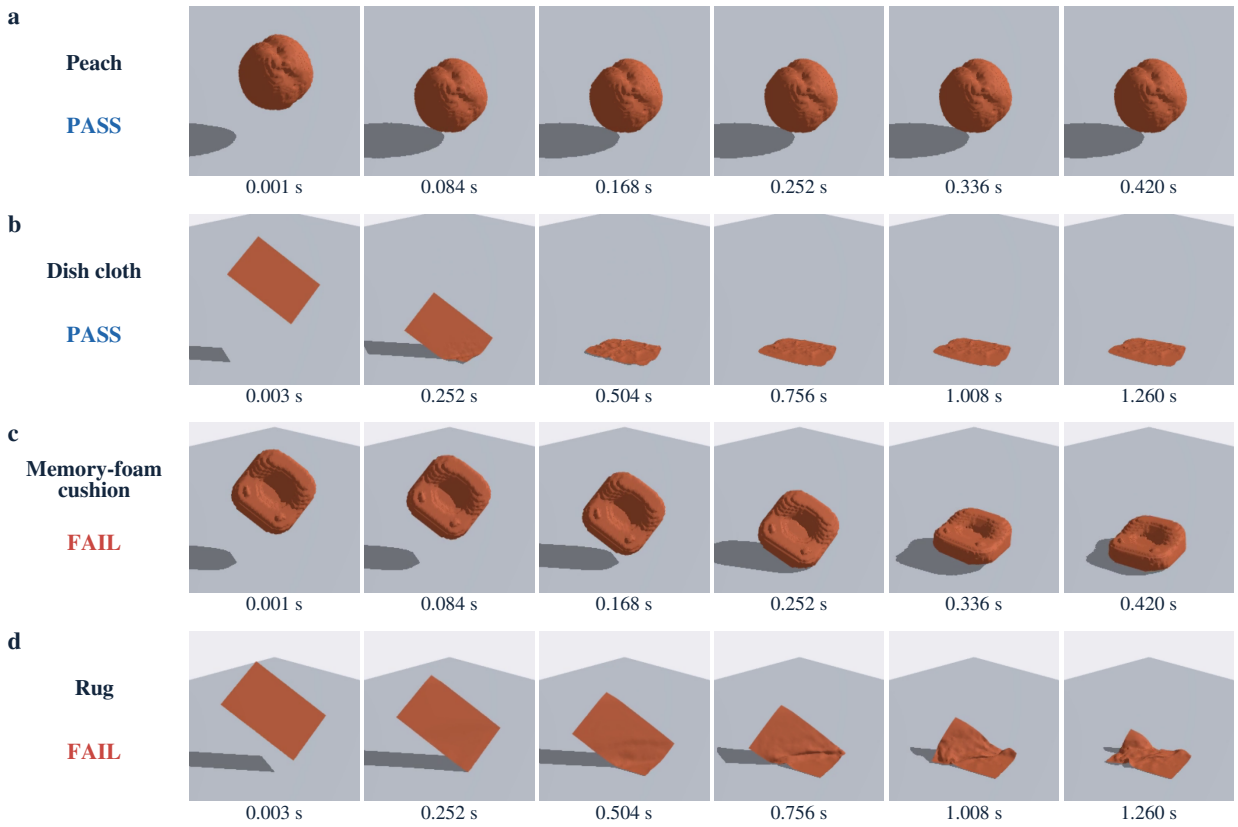


Figure 17 Drop tests: passing and non-settling cases. Rows (a, b) show passing peach and dish-cloth runs; rows (c, d) show memory-foam-cushion and rug runs that fail the recorded settling criterion. Frames are sampled at steps 1, 84, 168, 252, 336, and 420. Physical times use the effective solver step: 0.001 s in (a, c) and 0.003 s in (b, d). The first frame follows the first simulation step. The camera and crop remain fixed within each row; object scales differ across rows. The orange appearance comes from the original simulation videos.

G.2 Compression–release

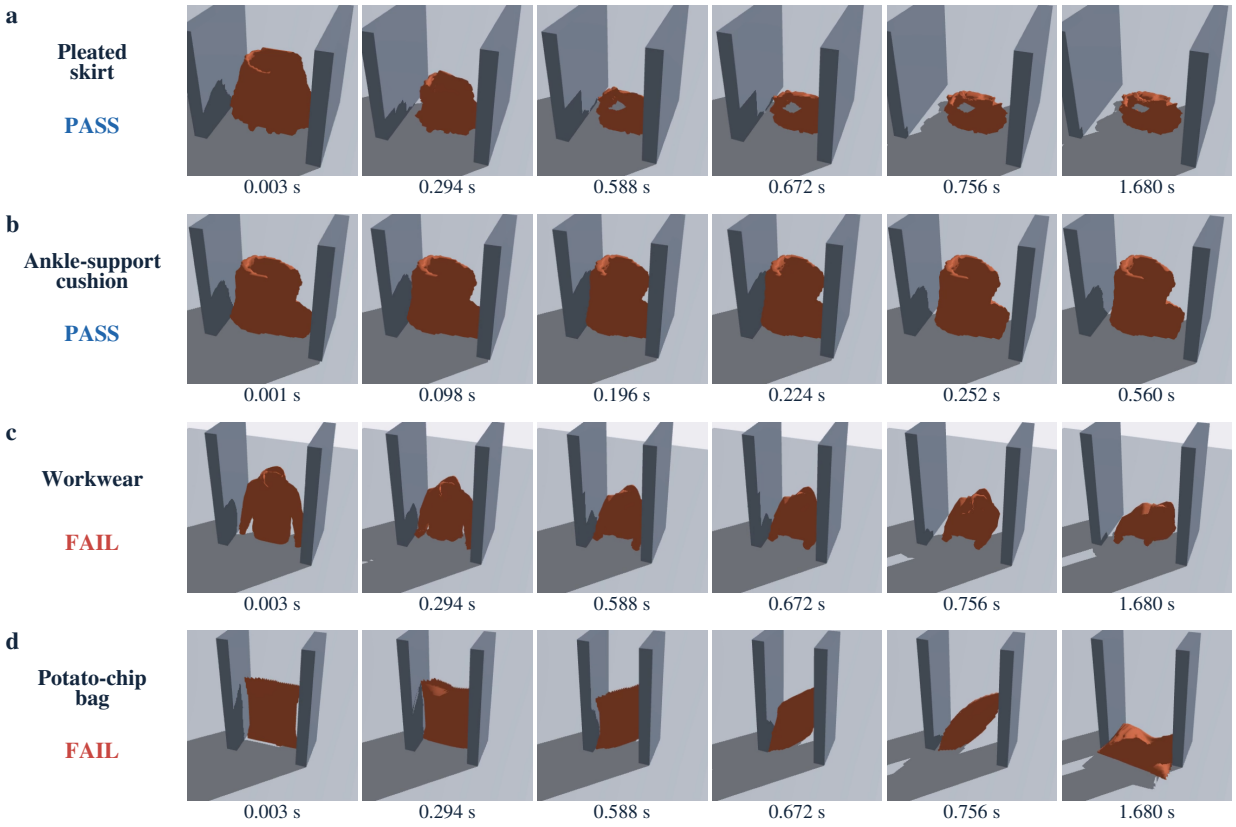


Figure 18 Compression–release tests: passing and non-settling cases. Rows (a, b) show passing pleated-skirt and ankle-support-cushion runs; rows (c, d) show workwear and potato-chip-bag runs that fail the recorded settling criterion. All four trajectories remain finite with no recorded ground penetration; FAIL denotes residual motion after release. Compression, holding, and retraction last 196, 28, and 28 steps, respectively. Frames show steps 1, 98, 196, 224, 252, and 560, using a 0.001 s solver step in (b) and 0.003 s in (a, c, d). Each row retains its fixed camera and full original frame; object scales differ across rows.