PHYSTRACE: A STANDARDIZED BENCHMARK FOR EVALUATING SIM-TO-REAL PHYSICAL FIDELITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical simulation provides scalable interaction environments for robot learning, yet understanding how physical fidelity affects the transfer of robot policies to the real world requires systematic investigation through comparable physical experiments, real-world observations, and task evaluations. We establish a benchmark for evaluating physical simulation quality with a focus on real-world transfer. Using six simulation platforms as experimental testbeds, the benchmark integrates fundamental physics experiments, everyday interaction tasks, and real-world experimental comparisons into a hierarchical evaluation framework. It establishes consistent experimental procedures for asset representation, physical properties, initial conditions, loading protocols, and observation specifications. Platform adaptation, native configuration verification, and temporal alignment of data ensure reproducible experimental conditions and traceable data pipelines, while explicitly documenting differences in physical implementations and parameter semantics across platforms. Through multifactor experimental matrices, controlled single-variable comparisons, and paired simulation–real-world comparisons, we evaluate the accuracy, stability, and parameter sensitivity of physical responses. We further use process data from control, contact, motion, and sensing to locate the stages at which discrepancies arise and analyze their propagation mechanisms. Under the tested conditions, agreement between simulation and real-world observations depends on the task, interaction stage, and configuration, with platforms exhibiting distinct responses during frictional transitions, transient collisions, and support transitions. Process observations and controlled comparisons provide evidence of how initial states, native numerical configurations, and contact responses contribute to some of these discrepancies. Building on these findings, supplementary short-horizon manipulation experiments with vision-language-action policies examine the relationships among physical discrepancies, task performance, and real-world transfer. The benchmark reveals that simulation fidelity is task- and stage-dependent: agreement in overall responses can coexist with discrepancies in critical interaction processes. This finding offers an important perspective on real-world transfer performance and supports a shift in simulation-based training for embodied models, from evaluating overall realism toward targeted calibration of task-critical physical processes.

1 INTRODUCTION

Physics simulation provides state transitions and interaction data for robot control, policy training, and pre-deployment validation. The same object, material label, and action sequence pass through geometry import, material pairing, collision detection, constraint solution, and time advancement on different platforms, ultimately producing observable motion and contact. Evaluation therefore requires a clear comparison object: what the common task specifies, what the platform executes, and which observations support correspondence with the real world.

We first organize elementary mechanics into controlled experiments, comparing equilibrium, onset, collision, and motion modes across six native implementations; we then extend these processes to household contact involving complex geometry, compliant materials, and multiple objects.

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.

PhysTrace[1] connects benchmark construction, standardized execution, same-task platform comparison, and source attribution. Its methodological contribution links comparison objects, common conditions, and process evidence, providing a consistent basis for comparing responses across six platforms and testing explanations for their differences.

2 Related work

Research on physics simulation and robot learning centers on multi-engine dynamics comparisons, real-reference validation, contact-model analysis, and manipulation-policy evaluation.

Engine comparisons and real-reference benchmarks. Erez et al. (2015) compare physics engines using common models, speed–accuracy relationships, energy, and momentum, finding that behavior varies with joint constraints, multibody scale, and contact processes. Collins et al. (2020) organize real manipulation records, protocols, and motion/force metrics into a reusable benchmark for specified robots and tasks; their analysis traces some object-motion differences to actuator response and contact arrival, with broader task and robot coverage as extensions. Acosta et al. (2022) identify contact parameters from real impacts and analyze sensitivity, showing that preferred parameters vary with system compliance and impact conditions, with edge/corner frictional impacts offering further modeling scope. Cloth validation extends physical references to deformation and material representation (Blanco-Mulero et al., 2024).

Contact models and interpretable numerical differences. Contact-model comparisons relate complementarity conditions, friction cones, compliance, and solver approximations (Le Lidec et al., 2024); unified rigid/deformable contact and simulation calibration offer numerical and parameter-level methods (Carpentier et al., 2024; Mehta et al., 2021). Factory and IndustReal link contact modeling to manipulation and transfer (Narang et al., 2022; Tang et al., 2023).

Manipulation scenes and learning benchmarks. SAPIEN and ManiSkill3 provide assets and manipulation tasks (Xiang et al., 2020; Tao et al., 2025); LIBERO and RoboCasa cover language, environment, and task generalization (Liu et al., 2023; Nasiriany et al., 2024). SIMPLER advances simulation-proxy evaluation (Li et al., 2025). SIMPLER treats visual and control differences as important evaluation factors and uses paired physical tests to verify correspondence in policy performance and behavior. Real reconstruction and online correction provide further transfer-data sources (Torne et al., 2024; Jiang et al., 2025).

Tactile and VLA applications. TACTO, Taxim, and DiffTactile support tactile simulation through imaging, example-based mapping, and contact physics, respectively (Wang et al., 2022; Si & Yuan, 2022; Si et al., 2024); their outputs serve interaction observations under the corresponding sensor and contact representations. Diffusion Policy, OpenVLA, and π_0 represent manipulation-policy and generalist-model approaches (Chi et al., 2023; Kim et al., 2025; Black et al., 2025). Inner Monologue illustrates the use of environmental feedback in language-model planning (Huang et al., 2023).

3 Evaluation framework and consistent configuration

The benchmark addresses three questions: whether platforms receive corresponding physical conditions, what responses these conditions produce, and where the response differences first arise. We connect common specifications, native adaptation, reference pairing, and process evaluation into a reusable workflow.

3.1 Comparison objects and reusable workflow

The comparison object is the state transition defined jointly by platform, physics backend, version, and configuration. Figure 1 shows the relation between experiments, observations, evaluation, and attribution.

[1]Project repository: `https://github.com/hoiyaipac4-lab/PhysTrace`.

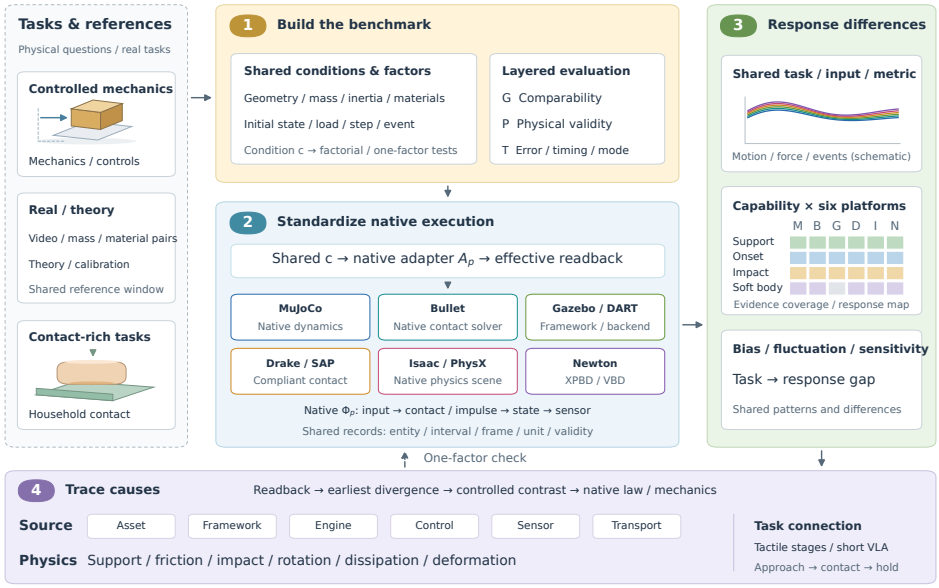

Figure 1: The unified benchmark: tasks and references define common conditions and G/P/T evaluation; native adapters map these conditions and read back effective settings; matched observations reveal response differences; controlled contrasts explain their sources. Source layers cross physical mechanisms. Solid arrows trace execution, the dashed arrow denotes a targeted control, and miniature curves/matrices illustrate output forms.

$$c = (\mathcal{G}, m, I, \mathcal{M}, x_0, u, h, n_s, K, \mathcal{E}), \qquad x_{k+1} = \Phi_p(x_k, u_k; A_p(c)). \tag{1}$$

Condition c specifies collision geometry $\mathcal{G}$, mass m, body-frame inertia I, pair-material parameters $\mathcal{M}$, initial pose/velocity x_0, force/torque or prescribed boundary motion u, outer step h, internal substeps n_s, native solver budget K, and event rules $\mathcal{E}$. Adapter A_p maps the common specification into native scene, material, and solver settings and retains requested and effective values. Native operator Φ_p advances one step under these effective settings, producing x_{k+1} for comparison.

Comparison qualification G checks entities, units, frames, inputs, time, and valid windows; physical validity P checks finite states, geometric contact, constraints, and energy; task response T characterizes trajectory accuracy, event timing, modes, and dissipation. Each layer retains its eligible denominator, defining the scope of trajectory, contact, and outcome comparisons.

3.2 Native observations, synchronization, and process evidence

Process observations comprise configuration/input, state/motion, interaction/constraints, solver process, sensing, and signal transmission, covering the chain from condition creation to evaluation. Records retain run/entity identity, source step, sampling phase, time interval, frame, units, and validity. The main association among these records is

$$u^{\text{applied}}_{[t_k, t_{k+1}]} \longrightarrow (\mathcal{C}, \mathbf{J})_{[t_k, t_{k+1}]} \longrightarrow x_{k+1} \longrightarrow y^{\text{sensor}}_{t_s}. \tag{2}$$

In Eq. (2), u^{applied} is the issued and read-back input; $\mathcal{C}$ stores interval contact pairs, points, and normals; $\mathbf{J}$ is the corresponding impulse; x_{k+1} is the end-of-interval state; and $y^{\text{sensor}}_{t_s}$ is the observation generated at source time t_s. Sensor observations can arrive after state sampling, so both source and reception times are retained.

Direct channels include native motion/joint states, available accelerations, contact/constraint responses, control readbacks, and raw sensor records. Contact-coordinate transformations and aggregates, image projection, errors, energies, and events are derived quantities. Raw high-frequency

Table 1: Evaluated platform/backend combinations and configuration provenance.

Platform/backend	Version	Native configuration tracked
MuJoCo	3.10.0	Regularization, friction cone, integration, stopping rules
PyBullet/Bullet	3.2.7	Sequential impulses, pair rules, iterations, speed limits
Gazebo Sim/DART	8.15.0/6.13.2	Components, constraint solver, input and sample phases
Drake/SAP	1.51.1	Point/hydroelastic contact, stiffness and relaxation
Isaac Sim/PhysX	4.5.0; 5.1*	Contact offsets; position/velocity budgets; deformables
Newton/XPBD	1.5.0; Warp 1.16	Substeps, iterations, correction and restitution phases

*Base rigid-body records use 4.5.0; deformable/appliance and scene renders use 5.1.

signals are stored independently, and phase summaries retain their windows and statistical definitions.

Synchronization first identifies pre-step, solve-interval, post-step, or asynchronous sensing phases, then pairs observations by source time. Video uses stream frame timestamps, and contact impulses accumulate over matching physical intervals.

3.3 Metrics and types of difference

On a fixed held-out observation set $\mathcal{H}$, trajectory, input, and event metrics are

$$E_x = \sqrt{\frac{1}{|\mathcal{H}|}\sum_{k\in\mathcal{H}} \|\Pi(x_k) - y_k^*\|^2}, \qquad e_u = \max_k \|u_k^{\text{read}} - u_k^{\text{ref}}\|, \tag{3}$$

$$\Delta t_e = t_e - t_e^*, \qquad A_{\text{mode}} = \frac{N_{\text{same}}}{N_{\text{eligible}}}. \tag{4}$$

Projection Π maps to registered image or metric coordinates, so E_x is measured in pixels or meters. Initial-state estimation and held-out evaluation use separate windows; the real probe supplies the input and the block region supplies independent motion observations. Failure rates use all attempts; mode agreement uses input-qualified complete configurations.

Bias is signed deviation from a reference; fluctuation is temporal variation within a physical phase; repeat dispersion is between-trial variation; sensitivity is response to a parameter change; anomalies follow finite-state, constraint, energy, or termination rules. Single-factor differences $\Delta Y = Y(c_{\xi_2}) - Y(c_{\xi_1})$ are interpreted with native channels.

3.4 Consistent configuration

Consistent configuration covers assets, inputs, material pairs, and time semantics. Standardization implements common specifications, native-parameter semantic mapping, effective-configuration readback, and matched observation windows, establishing corresponding conditions for the same physical task across six platforms. Platform selection covers complementary native implementations; configuration checks align the physical question while retaining contact representations, constitutive models, and solver differences as measured properties.

Table 1 lists the tested combinations. SI units, gravity, object identity, collision shape, mass/inertia, input trajectories, and recording periods are shared; visual and collision geometry are stored separately. Material enters through effective pairs $\mu_{AB} = C_p(\mu_A, \mu_B)$, while contact offsets, stiffness, damping, restitution, and combination rules retain native meanings.

The main rigid-body matrix uses 60, 120, and 240 Hz, with $h = 1/f$. Physical step and solver-iteration budget are separate variables: the former changes geometry/contact update times, while the latter acts on the within-step constraint solution (Macklin et al., 2019). Control, rendering, and recording each have independent clocks.

Tissue packages, soft bread, and cushions use common coarse volumetric meshes, measured mass/table friction, and shared elastic priors for native soft-body fall comparisons in MuJoCo, Bullet,

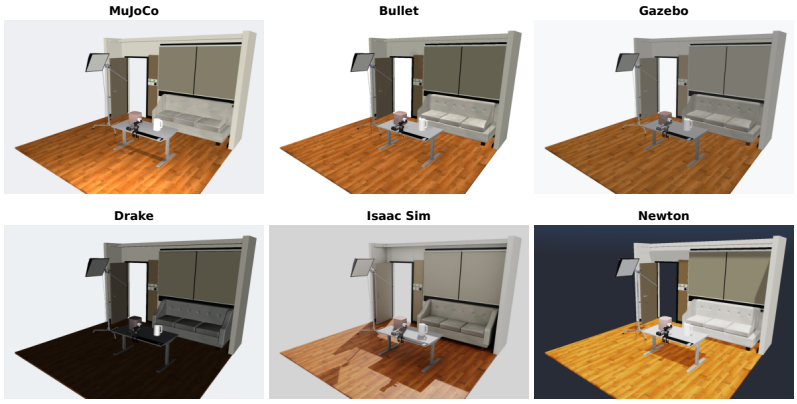

Figure 2: Common scene assets in six platforms.

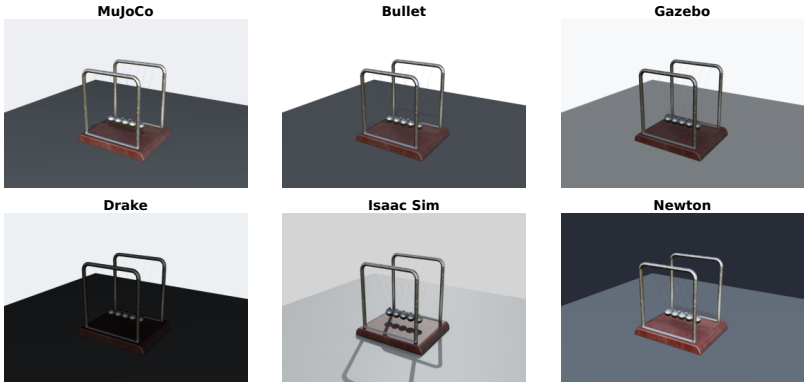

Figure 3: Complete existing cradle asset in six native renderers. The original base, textured frame, balls, eyelets, and paired thin rods share the same rest geometry and camera. The source has 18 rigid bodies and 22 joints; these images document visual assembly. Figure 5 separately identifies the controlled ideal-hinge dynamics used for mechanism analysis.

Drake, Newton, and Isaac; Gazebo records the volumetric-soft-body adaptation scope of the verified interface.

Adapters preserve joint axes, local anchors, and transformed centers of mass. Configuration readback checks the native realization of these quantities, while channel mapping preserves force, impulse, correction, and reaction-wrench semantics. Figures 2 and 3 document shared visual assemblies; dynamic comparisons retain their registered collision and constraint models.

4 Real references and experimental design

Assets cover basic geometries with different materials, masses, shapes, and scales, together with household objects featuring complex surfaces, mass distributions, and compliance. Contact surfaces and placement/loading arrangements provide controlled environmental variation. Weighing and force measurements supply mass and effective material-pair properties; direct measurements or image calibration supply dimensions and their ranges. The six platforms share these asset identities, physical targets, and release/loading conditions, while their native backends generate motion. Exact masses, candidate dimensions, friction loads, and pendulum parameters are retained in Appendix A.

Friction uses effective material pairs under the specified force-measurement protocol, estimated dimensions use registered ranges, and soft-body elasticity uses explicit material priors. These correspond to measured inputs, geometric sensitivity, and model assumptions. The main text compares

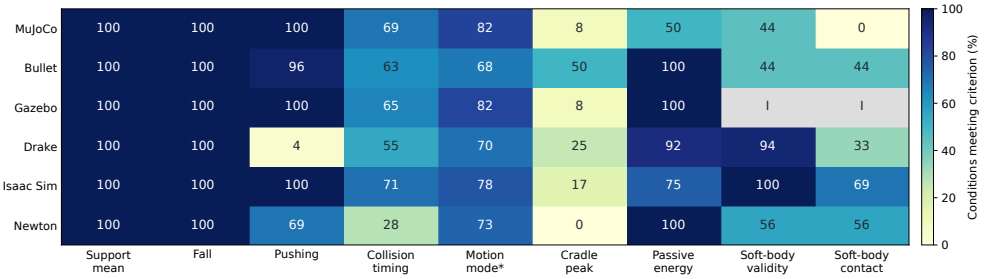

Figure 4: Task-conditioned response map. Cells show percentages satisfying the declared criterion. Support uses steady means, fall uses same-initial-state analytic motion, and pushing/timing/mode/peak use real references. Energy and soft-body columns apply physical/numerical screens. Mode* uses input-qualified complete subsets. Gray denotes the evaluated interface's coverage.

response patterns under common inputs; the appendix reports parameters, mechanical assumptions, repetition units, and observation scope, allowing overall findings to be traced to specific conditions.

Reference reconstruction preserves video timestamps, object identity, calibration features, and visibility. Initial-state estimation and later evaluation use separate windows; prescribed-boundary and object-motion observations have distinct roles. Scale and initial-state sensitivity use the same held-out frames, connecting geometric uncertainty to the evaluated response.

Repeated software restarts quantify numerical replay, while independent real operations characterize trial variability. Conditions sharing a real recording retain that dependence. Each comparison reports its sampling unit, usable interval, and terminal outcome; Appendix A provides the corresponding denominators and rules. This separates parameter sensitivity from between-trial dispersion and gives each physical reference an explicit scope.

5 Simulation results and mechanism attribution

5.1 Overall response and representative processes

Three connected patterns emerge: total support and pre-contact motion show substantial commonality; contact establishment, tangential torque, and mode switching enlarge response differences; coupled transmission and soft-body support further amplify temporal-discretization and representation effects.

Within each platform-metric combination in Figure 4, conditions have equal weight, and the proportion satisfying the criterion is mapped to a 0–100 scale. Thresholds and parameter uncertainty follow Appendix A.5; columns address distinct physical questions.

All six platforms satisfy the 1% support-mean threshold in 18 conditions each, evaluated over registered steady windows and three restarts per condition.

All 270 fall conditions complete; held-out median RMSE is 20.43/9.50/6.81 px at 60/120/240 Hz. The common offset follows the semi-implicit update

$$v_{k+1} = v_k - gh, \quad z_{k+1} = z_k + hv_{k+1}, \qquad z_k - z_{\text{exact}}(kh) = -\tfrac{1}{2}gh(kh). \tag{5}$$

Same-step six-platform position spread is submicrometer, while interpolation using stored timestamps yields a few micrometers. Both differ from the common millimeter-scale integration residual, explaining nearly overlapping raw trajectories.

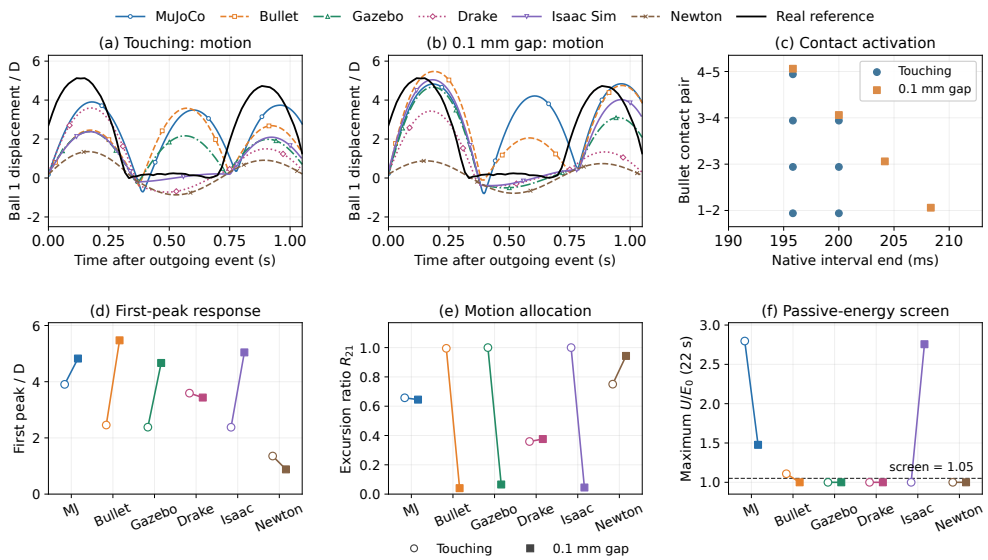

Figure 5: Evidence chain for the 240 Hz ideal-hinge controls. (a,b) Output-ball motion for touching and 0.1 mm gap, aligned to each outgoing event; the black video reference has a reconstructed release angle of about $51°$, compared with the simulated release angle of $60°$. (c) Absolute native Bullet contact timestamps. (d) Paired first peaks measure the macroscopic change. (e) Second-to-first ball excursion ratio R_{21} over 0.35 s identifies motion allocation. (f) Maximum gravitational potential divided by initial energy over 22 s, with a 1.05 screening line, adds an independent passive-energy check.

5.2 Driven contact: onset, pushing, and mode transitions

For real pushing at 240 Hz with friction 0.20 on both pairs, the median position errors across eight clips are 2.20/2.15/2.15/2.14 mm for MuJoCo/Bullet/Gazebo/Isaac and 43.58 mm for Drake. Newton's median is 2.58 mm across seven completed clips, with the remaining clip recorded separately.

The representative condition reaches 13.15 mm lateral displacement and $-39.18°$ yaw at 1.5 s; boundary position error across the 72 Drake runs is at most 42.62 μm.

Of 40 real sliding–tipping conditions, 27 slide and 13 tip; Bullet agrees in 19/40, 31/40, and 32/40 at 60/120/240 Hz, while MuJoCo agrees in 33/40 at each frequency. Pair-force and application-height controls connect loading to final mode (Appendix B).

5.3 Newton's cradle: from contact topology to motion allocation

Newton's cradle concentrates the interaction-pathway finding into a controllable chain: prescribed release, gap-only variation, contact establishment, and ball-wise transmission.

Figure 5 compares zero/0.1 mm gaps in the registered five-planar-hinge model at 240 Hz and $60°$ release: initial positions agree within 15 nm, with one deterministic run per configuration (Appendix B).

The gap increases first peaks in MuJoCo/Bullet/Gazebo/Isaac and decreases them in Drake/Newton. Over the first 0.35 s outgoing window, the second-to-first ball excursion ratio falls from approximately one to 0.040/0.066/0.045 in Bullet/Gazebo/Isaac. Release angles differ; the within-platform pair isolates the gap effect.

Over 22 s, gap-run maximum potential/initial-energy ratios reach 1.477 and 2.756 for MuJoCo and Isaac, exceeding the 1.05 passive-energy screen; Bullet changes from 1.109 to 1.000. Peak correspondence, motion allocation, and long-window validity therefore carry distinct evidence. Section 5.5 connects these responses to contact activation; identifying each energy deviation's numerical source requires the corresponding solver evidence.

5.4 Household objects and compliant contact

Node positions, deformation-gradient determinants, and contact distances jointly assess soft-body physical validity, separating finite deformation from contact support (Figure 4).

Common appliance opening-input replays yield similar tracking. In the independent eccentric-hinge test, Newton XPBD's linear/angular relaxation pair changes response substantially, while iteration sweeps leave the approximately 47.7% endpoint deviation nearly unchanged. Matched inertia, input, and analytic recurrence support this configuration-specific effect. Real hand signals supply interaction phases in raw units; native joint effort retains its generalized-force units and independent dynamic reference (Appendix D).

5.5 Full-process attribution and platform capabilities

Results follow two intersecting axes: the system layer identifies where a difference enters, and physical capability identifies what it changes.

Attribution spans framework, engine, asset, controller, sensor, and signal/communication layers, crossed with support, friction, impulse, rotation, and dissipation mechanisms. A strong evidence chain combines input readback, time-ordered native intermediate variables, controlled parameter contrasts, and native algorithms or independent physical balance.

Effective-input checks distinguish adapter realization from downstream dynamics. The pushing contrast fixes floor friction at 0.20 and sets only probe friction to zero: MuJoCo/Bullet/Isaac probe tangential forces vanish, while c3/c5 endpoint changes remain within 0.43 mm. Trajectory agreement and force-process agreement therefore require their own observations.

Bullet's native records resolve the cradle mechanism. With touching geometry, all four adjacent pairs carry positive normal force in the step ending at 195.833 ms. A 0.1 mm gap spreads activation across successive steps from 195.833 to 208.333 ms. An isolated two-ball control gives an effective relative-velocity restitution of 0.959, close to the registered 0.95 pair target. These observations connect geometry to simultaneous versus sequential activation and then to macroscopic motion allocation. Other platforms' mechanism statements remain within the scope of their available channels.

First, free motion and total equilibrium provide common baselines, with temporal discretization and observation reconstruction explaining separate error sources. Second, contact establishment and mode transitions amplify material pairing, geometry, and local torque; close endpoints can still follow different force histories. Third, coupled contacts and compliant support combine constraint timing, constitutive behavior, and spatial discretization, making response dependent on configuration and phase. Full-process attribution thus converts platform differences into identifiable, testable mechanism conditions, supporting model and numerical choices for the contact stages a task requires.

The URDF–scan comparison adds an asset-layer diagnostic: millimeter-scale surface mismatch and local installation offsets are evaluated before attributing task discrepancies to the solver.

6 VLA learning and cross-platform execution

The physical benchmark shows that similar trajectories can accompany different contact processes; learning evaluation therefore distinguishes training source, execution platform, and task phase. Training sources change demonstration state–action distributions, while execution platforms change closed-loop state advancement and tracking; both influence subsequent observations and actions.

The source-training protocol uses common joint definitions, image dimensions, and action channels, with fixed within-source validation identities, 3,000 updates per source, and one training seed. Its process matrix crosses four sources, six execution platforms, and three tasks at a fixed seed. Common-URDF forward kinematics evaluates commanded paths, while native end-effector states evaluate tracking (Appendix F). These paired views separate the action presented to the platform from its physical realization.

Across 72 episodes, actual-to-command position error is 1.77–4.28 mm and command-to-nominal-path error is 96.30–129.17 mm; registered approach events are zero, and the 60 episodes with ad-

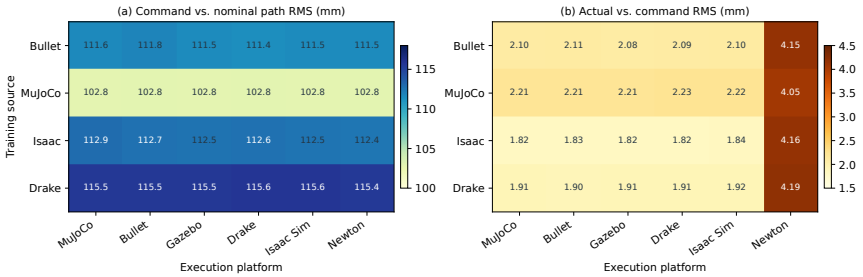

Figure 6: Source-model × execution-platform process comparison. Each cell averages push, pick, and place with a fixed seed. A: commanded FK path relative to a common nominal path; B: actual end-effector path relative to its command. Independent color scales and millimeter units distinguish action generation from execution tracking.

equate active-contact coverage contain zero contact events. The joint evidence locates the main separation in the approach path, followed by platform tracking differences.

An independent action-representation diagnosis pairs sampling paths for known placeholder dimensions at the same checkpoint across three tasks and three seeds. The constrained path reaches sampled contact in 8/9 episodes, with full-task success of 1/9 versus 0/9; intervals and matched initial/noise conditions are detailed in Appendix F.

A separate same-weight six-platform grasping study compares fixed text with stage-semantic feedback over 48 episodes: success is 0/48 and all returned robot–object contact records are zero. The two modes retain model, action protocol, and sampling identities separate from the source-training matrix; implementation and detailed contrasts appear in Appendix F.

Source models affect command timing and paths, while execution backends affect tracking; paired action-representation diagnostics with matched initial states and noise further change contact arrival. Simulation quality should therefore be interpreted at the physical stages a policy actually reaches: approach first tests action/control consistency, and subsequent contact tests friction, restitution, and support propagation.

7 Conclusion

PhysTrace links standardized six-platform experiments to traceable process evidence. Under the tested conditions, physical fidelity is task- and stage-dependent: common support and free-motion responses coexist with differences in contact establishment, motion modes, and coupled transmission. Similar trajectories can conceal different force histories. Cradle gap controls, contact timestamps, and pair tests connect activation to motion allocation; independent hinge mechanics and relaxation sweeps identify a configuration-specific dynamic effect. These findings guide targeted calibration of task-critical processes. VLA records locate the dominant observed separation before contact, prioritizing action/control consistency in those cohorts. This establishes a process-level basis for the asset-validation and real-transfer tests outlined next.

8 Future work

Building on the shared asset specifications, physical calibration, and URDF–scan diagnostics established here, we will develop a pipeline that converts videos, images, or 3D scans into simulation-ready assets. The benchmark will guide cross-platform validation of reconstructed geometry, articulation, and physical properties. Expanding object and material coverage will support controlled VLA training comparisons using these assets. With matched policies, training budgets, and action interfaces, we will test how benchmark-guided asset calibration affects contact arrival, force and motion histories, and real-world transfer. This connects asset generation to the paper's central principle: shared physical conditions and process evidence make simulation differences interpretable and their learning implications testable.

AI ASSISTANCE DISCLOSURE

Generative AI tools were used to aid and polish writing, including translation, language revision, and organization of author-directed manuscript content; to retrieve and discover relevant literature and technical documentation; and to support research ideation and execution, including experimental planning, implementation and debugging of software, data organization, analysis, and figure preparation. AI-assisted suggestions informed the workflow under author direction. The authors are responsible for verifying the cited sources, experimental records, code, analyses, and interpretations, and for the accuracy and integrity of the final manuscript.

REFERENCES

Brian Acosta, William Yang, and Michael Posa. Validating Robotics Simulators on Real-World Impacts. *IEEE Robotics and Automation Letters*, 7(3):6471–6478, 2022. doi: 10.1109/lra.2022.3174367. URL https://doi.org/10.1109/lra.2022.3174367.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.010. URL https://www.roboticsproceedings.org/rss21/p010.html.

David Blanco-Mulero, Oriol Barbany, Gokhan Alcan, Adrià Colomé, Carme Torras, and Ville Kyrki. Benchmarking the Sim-to-Real Gap in Cloth Manipulation. *IEEE Robotics and Automation Letters*, 9(3):2981–2988, 2024. doi: 10.1109/lra.2024.3360814. URL https://doi.org/10.1109/lra.2024.3360814.

Justin Carpentier, Quentin Le Lidec, and Louis Montaut. From Compliant to Rigid Contact Simulation: a Unified and Efficient Approach. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.108. URL https://www.roboticsproceedings.org/rss20/p108.html.

Alejandro Castro, Frank Permenter, and Xuchen Han. An unconstrained convex formulation of compliant contact. *IEEE Transactions on Robotics*, 39(2):1301–1320, 2023. doi: 10.1109/tro.2022.3209077. URL https://arxiv.org/abs/2110.10107.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://roboticsproceedings.org/rss19/p026.html.

Jack Collins, Jessie McVicar, David Wedlock, Ross Brown, David Howard, and Jürgen Leitner. Benchmarking Simulated Robotic Manipulation through a Real World Dataset. *IEEE Robotics and Automation Letters*, 5(1):250–257, 2020. doi: 10.1109/lra.2019.2953663. URL https://arxiv.org/abs/1911.01557.

Erwin Coumans and Yunfei Bai. *PyBullet Quickstart Guide*, n.d. URL https://github.com/bulletphysics/bullet3/blob/master/docs/pybullet_quickstart_guide/PyBulletQuickstartGuide.md.html. Official Bullet repository guide; accessed 2026-09-26.

DART Developers. DART v6.13.2: Boxed lcp constraint solver, n.d. URL https://github.com/dartsim/dart/blob/v6.13.2/dart/constraint/BoxedLcpConstraintSolver.cpp. Version-pinned native constraint assembly and solver implementation; accessed September 2026.

Drake Developers. *Hydroelastic Contact User Guide*, n.d. URL https://github.com/RobotLocomotion/drake/blob/v1.51.1/multibody/hydroelastics/hydroelastic_user_guide_doxygen.h. Drake v1.51.1 documentation source; accessed 2026-09-17.

Tom Erez, Yuval Tassa, and Emanuel Todorov. Simulation Tools for Model-Based Robotics: Comparison of Bullet, Havok, MuJoCo, ODE and PhysX. In *ICRA*, pp. 4397–4404, 2015. doi: 10.1109/icra.2015.7139807. URL https://homes.cs.washington.edu/~todorov/papers/ErezICRA15.pdf.

Gazebo Developers. *Gazebo Sim 8: Physics Engines*, n.d. URL https://gazebosim.org/api/sim/8/physics.html. Major-version 8 documentation; accessed 2026-09-17.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1769–1782. PMLR, 2023. URL https://proceedings.mlr.press/v205/huang23c.html. Conference held in 2022; proceedings published in 2023.

Yunfan Jiang, Chen Wang, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. TRANSIC: Sim-to-Real Policy Transfer by Learning from Online Correction. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1691–1729. PMLR, 2025. URL https://proceedings.mlr.press/v270/jiang25a.html. Conference held in 2024; proceedings published in 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025. URL https://proceedings.mlr.press/v270/kim25c.html. Conference held in 2024; proceedings published in 2025.

Quentin Le Lidec, Wilson Jallet, Louis Montaut, Ivan Laptev, Cordelia Schmid, and Justin Carpentier. Contact Models in Robotics: A Comparative Analysis. *IEEE Transactions on Robotics*, 40:3716–3733, 2024. doi: 10.1109/tro.2024.3434208. URL https://simple-robotics.github.io/publications/contact-models/.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating Real-World Robot Manipulation Policies in Simulation. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3705–3728. PMLR, 2025. URL https://proceedings.mlr.press/v270/li25c.html. Conference held in 2024; proceedings published in 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *NeurIPS Datasets and Benchmarks*, pp. 44776–44791, 2023. URL https://proceedings.nips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Miles Macklin, Matthias Müller, and Nuttapong Chentanez. XPBD: Position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, pp. 49–54, 2016. doi: 10.1145/2994258.2994272. URL https://mmacklin.com/xpbd.pdf.

Miles Macklin, Kier Storey, Michelle Lu, Pierre Terdiman, Nuttapong Chentanez, Stefan Jeschke, and Matthias Müller. Small steps in physics simulation. In *Eurographics/ACM SIGGRAPH Symposium on Computer Animation*, pp. 1–7. ACM, 2019. doi: 10.1145/3309486.3340247. URL https://mmacklin.com/smallsteps.pdf.

Bhairav Mehta, Ankur Handa, Dieter Fox, and Fabio Ramos. A User's Guide to Calibrating Robotic Simulators. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 1326–1340. PMLR, 2021. URL https://proceedings.mlr.press/v155/mehta21a.html. Conference held in 2020; proceedings published in 2021.

MuJoCo Developers. *MuJoCo Computation: Numerical Integration*, n.d. URL https://mujoco.readthedocs.io/en/3.10.0/computation/index.html. Version 3.10.0; accessed 2026-09-26.

Yashraj Narang, Kier Storey, Iretiayo Akinola, Miles Macklin, Philipp Reist, Lukasz Wawrzyniak, Yunrong Guo, Adam Moravanszky, Gavriel State, Michelle Lu, Ankur Handa, and Dieter Fox. Factory: Fast Contact for Robotic Assembly. In *Proceedings of Robotics: Science and Systems*, 2022. doi: 10.15607/RSS.2022.XVIII.035. URL https://www.roboticsproceedings.org/rss18/p035.html.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-Scale Simulation of Household Tasks for Generalist Robots. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.050. URL https://www.roboticsproceedings.org/rss20/p050.html.

Newton Developers. Newton v1.5.0 native rigid-body integration, n.d.a. URL https://github.com/newton-physics/newton/blob/v1.5.0/newton/_src/solvers/solver.py. Installed source lines 92–100 archived with the rotation diagnostic; accessed 2026-09-19.

Newton Developers. Newton v1.5.0: SolverXPBD implementation and documentation, n.d.b. URL https://github.com/newton-physics/newton/blob/v1.5.0/newton/_src/solvers/xpbd/solver_xpbd.py. Tag v1.5.0; accessed 2026-09-17.

NVIDIA. *PhysX 5.4.1 Documentation: Rigid Body Dynamics*, n.d. URL https://nvidia-omniverse.github.io/PhysX/physx/5.4.1/docs/RigidBodyDynamics.html. Algorithm documentation, version 5.4.1; accessed 2026-09-17.

Zilin Si and Wenzhen Yuan. Taxim: An Example-Based Simulation Model for GelSight Tactile Sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022. doi: 10.1109/lra.2022.3142412. URL https://arxiv.org/abs/2109.04027.

Zilin Si, Gu Zhang, Qingwei Ben, Branden Romero, Zhou Xian, Chao Liu, and Chuang Gan. DIFFTACTILE: A Physics-based Differentiable Tactile Simulator for Contact-rich Robotic Manipulation. In *ICLR*, pp. 7164–7183, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1cc70be9fb6a83bc46cf4ac21a91e0b0-Abstract-Conference.html.

Bingjie Tang, Michael A Lin, Iretiayo A Akinola, Ankur Handa, Gaurav S Sukhatme, Fabio Ramos, Dieter Fox, and Yashraj S Narang. IndustReal: Transferring Contact-Rich Assembly Tasks from Simulation to Reality. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.039. URL https://www.roboticsproceedings.org/rss19/p039.html.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-Kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Demonstrating GPU Parallelized Robot Simulation and Rendering for Generalizable Embodied AI with ManiSkill3. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.021. URL https://www.roboticsproceedings.org/rss21/p021.html.

Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling Reality through Simulation: A Real-to-Sim-to-Real Approach for Robust Manipulation. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.015. URL https://www.roboticsproceedings.org/rss20/p015.html.

Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. TACTO: A Fast, Flexible, and Open-source Simulator for High-Resolution Vision-based Tactile Sensors. *IEEE Robotics and Automation Letters*, 7(2):3930–3937, 2022. doi: 10.1109/lra.2022.3146945. URL https://arxiv.org/abs/2012.08456.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A SimulAted Part-Based Interactive ENvironment. In *CVPR*, pp. 11097–11107, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Xiang_SAPIEN_A_SimulAted_Part-Based_Interactive_ENvironment_CVPR_2020_paper.html.

A Experimental protocols, assets, and statistical scope

A.1 From experiment design to attribution

Each condition follows seven steps. (1) State the physical question and distinguish prescribed inputs from evaluated outputs. Select analytic, real-motion, or sensor references with explicit assumptions. (2) Register object identity, visual/collision geometry, dimensions, mass, center of mass, inertia, and material pairs. Tag measured values, engineering values, image-estimated ranges, and priors. (3) Select factors from the governing mechanics and distinguish parameter conditions, real trials, and software restarts. (4) Map common inputs into native APIs and read back effective coordinates, inertia, initial state, contact settings, and timing. (5) Execute native dynamics and record inputs, contact/constraints, state, and outcome. (6) Compare matching intervals and retain all attempted outcomes in the appropriate denominator. (7) identify the earliest separating channel and test the candidate with controlled parameters and independent mechanics.

This sequence specializes to probe tracking, block yaw, separation, and displacement plateau in pushing; to common inertia/torque, native constraint response, and angle in the hinge test; and to command generation, execution tracking, and contact access in policy evaluation. It preserves one shared evidence logic while allowing different instruments and native outputs.

A.2 Time protocols and sample units

The principal rigid-body matrix uses 60/120/240 Hz. The cradle's thin-rod angle–gap matrix uses 960 Hz, with separate frequency controls; its ideal-hinge micro-gap case uses 240 Hz. Native camera/state rendering is separate from advancement. Single soft bodies run for 1 s at 240/960/3840 Hz; stacks run for 0.7 s at 3840 Hz. Policy-source experiments use 960 Hz integration, 240 Hz state recording, and 15 Hz actions. Each channel identifies whether its frequency refers to advancement, sensing, or storage.

The controlled rigid-body matrix contains 1,944 executions. Each of six tasks has six conditions, three frequencies, and three software restarts per platform. Among 546 valid condition cells, 458 have zero restart standard deviation, expressing numerical reproducibility in the registered environment. The physical-video inventory has 157 clips and 122 principal rigid-body clips. Most rigid-body physical conditions have one recording; real tactile conditions have ten independent operations. The soft-body collection has 225 single-body runs (180 base, 36 iteration controls, nine contact controls), 30 stack runs, and ten friction controls. Each exact soft-body configuration has one run. These sampling units separately support deterministic sensitivity, independent-operation variability, and real nominal references.

Table S1: Real-reference reproduction queues and outcome accounting. Prefix analysis and complete-window evaluation retain separate denominators.

Queue	Attempts	Complete	Scope
Fall	270	270	15 clips, six platforms, three frequencies; 443 held-out frames
Pushing	432	430	Eight clips, three friction values, three frequencies, six platforms
Sliding–collision	450	417	25 clips; 3 Drake and 30 Newton nonfinite outcomes
Sliding–tipping	720	692	40 clips; 674 complete, input-qualified conditions
Cradle principal controls	96	96	72 release/gap configurations and 24 time/angle controls
Release symmetry	6	6	One equal-angle two-ball condition per platform
Restitution mechanisms	6	6	Three local Newton and three independent PhysX controls
Soft bodies	225	Per criterion	Finite state, orientation, and penetration assessed separately
Soft stacks and friction	40	40	Five available native backends

A.3 Condition factors, loading, and termination

R00–R05 are local indices reused within each controlled task. They enumerate the Cartesian product of two declared factors: support mass 0.043/0.1314 kg and support-spacing scale 0.7/1.0/1.3; horizontal onset friction 0.2/0.5 and force-ramp rate 0.1/0.2/0.4 mg/s; head-on impact speed 0.25/1.5 m/s and nominal restitution target 0.2/0.65/0.9; fall/rebound height 0.1/0.5 m and sphere radius 4/7.9365/16 mm; sliding–tipping aspect ratio 1/3 and loading height 0.3/0.6/0.9 of object height. The incline archive includes friction 0.2/0.5 and 4/12/24 mm geometric contact controls. The selected material slice fixes 4 mm and has 54 common event observations across three platforms, two friction values, three frequencies, and three restarts.

Incline loading begins after 0.5 s settling and increases at $2.5°$/s until sustained motion or the registered $80°$ bound. The base event uses 3 mm displacement, 0.03 m/s relative speed, and 0.08 s persistence. The separate real a1 reproduction uses a 4 mm threshold. Relative downslope motion is

$$v_{\text{rel}} = \mathbf{t}^{\top}\left[\mathbf{v}_b - \mathbf{v}_o - \boldsymbol{\omega}_{\text{ramp}} \times (\mathbf{r}_b - \mathbf{r}_o)\right].$$

World motion and ramp-relative motion therefore have distinct meanings. Rotation and contact identify the applicability of a translational onset event. For horizontal force ramps, confirmation delay contributes $\Delta F = \dot{F}\Delta t$ and discrete force resolution is $\dot{F}h$.

Real pushing c1–c8 covers four block materials in cubic and elongated forms. Short elongated-block candidate masses are 2.5/16.6/46.2/50.6 g. Probe image samples every 12 frames define loading-boundary knots; independent block features every four frames define evaluation. Effective common friction values 0.10/0.20/0.30, three frequencies, and six platforms form 432 conditions. Drake's probe controller uses $k_p = 4 \times 10^5$ N/m and $k_d = 100$ N s/m; all 72 input paths meet a 0.1 mm qualification threshold. Pair-isolation tests keep floor friction fixed and change probe friction only.

Sliding–tipping uses 40 real conditions: 27 slide and 13 tip. Candidate elongated geometry is $10 \times 20 \times 50$ mm. Replayed probe trajectories retain their estimated height; the largest observed vertical drift is 7.586 mm. A native tipping event requires the body vertical axis to reach $60°$ for 25 ms; sustained $10°$ tilt is recorded separately. The sliding–collision set has 1,251 reviewed positions in 25 clips: 200 define initial conditions and 1,051 are held out. Ramp angle, seam geometry, and weak-perspective projection belong to each reconstruction record.

A.4 Reference reconstruction and uncertainty

Processing retains the video stream time base, object identity, contact-surface identity, and calibration feature. Visible dimensions define a metric conversion or bounded candidate scale. Initial-state frames determine release pose/velocity; later frames define a fixed evaluation window. Probe and block regions are independent roles in pushing. Missing or occluded features carry availability flags. Raw time, event-relative time, and interpolation rules remain separately recoverable. Scale and initial-state sensitivity are tested on the same held-out frames.

Repeated restarts quantify deterministic replay; single real releases define nominal trajectory differences and parameter ranges. Bootstrap or confidence intervals apply only to their declared sampling unit. Tactile operations support independent-trial summaries; parameter-grid cells sharing one real clip retain that dependence. Failure rate uses all attempted runs. A prefix trajectory score describes its usable interval and is accompanied by the terminal outcome. The study's screening tolerances organize retrospective interpretation of the recorded conditions.

A.5 Response-map criteria and integer denominators

The score in main Figure 4 is $S_{pj} = 100 \sum_{c \in \mathcal{C}_{pj}} \mathbf{1}\{c \text{ satisfies criterion } j\}/|\mathcal{C}_{pj}|$. One condition contributes one count. The thresholds express separate questions: (i) support mean within 1% of mg; (ii) same-initial-state fall RMSE within $\max(2\text{ mm}, 0.05L)$ of analytic motion; (iii) complete qualified pushing RMSE within $\max(2\text{ mm}, 0.05L)$ of real feature displacement; (iv) sliding–collision first-contact interval within 20 ms of the real interval; (v) mode agreement among complete input-qualified conditions; (vi) cradle first peak within 10% of the real peak; (vii) maximum registered mechanical energy/initial energy at most 1.05; (viii) complete finite, positively oriented soft bodies with penetration below 50 mm; and (ix) valid soft-body contact penetration at most 2 mm. L denotes the corresponding trajectory's displacement scale.

Table S2: Passing/evaluated conditions, in the same column order as the main response map. Mode uses a conditional subset; soft-body coverage follows the tested native interfaces.

	Support	Fall	Push	Timing	Mode	Peak	Energy	Soft valid	Soft contact
MuJoCo	18/18	45/45	72/72	52/75	99/120	1/12	6/12	16/36	0/36
Bullet	18/18	45/45	69/72	47/75	82/120	6/12	12/12	16/36	16/36
Gazebo	18/18	45/45	72/72	49/75	98/120	1/12	12/12	–	–
Drake	18/18	45/45	3/72	41/75	71/101	3/12	11/12	34/36	12/36
Isaac	18/18	45/45	72/72	53/75	93/120	2/12	9/12	36/36	25/36
Newton	18/18	45/45	50/72	21/75	68/93	0/12	12/12	20/36	20/36

The support map combines three restart means per condition, using the registered last 1 s for MuJoCo/Isaac/Gazebo and last 0.5 s for Bullet/Drake/Newton. This mean-screening question is distinct from common-window force fluctuation. The fall column measures native integration relative to a same-initial-state analytic reference; real-image residuals are reported separately. The cradle energy screen describes excess energy, while envelope tests describe dissipation. Soft-body validity and contact response have separate thresholds. These distinctions preserve the physical interpretation of each column.

A.6 URDF geometry and real scans: layered asset validation

The Piper study compares the articulated engineering model, its visual/collision assets, and three whole-arm OBJ scans: two enabled poses and one disabled pose. Scan coordinates are interpreted in millimeters and mapped to SI units by 0.001 scaling plus a registered rigid transform. Whole-arm scans have a single unsegmented surface; the fourth supplied part is a G-shaped screw clamp. The analysis preserves an articulated URDF skeleton and evaluates local surfaces against the appropriate link. It distinguishes scan-shape discrepancy, installation geometry, estimated pose, and dynamic parameters.

The supplied comparison report records signed mean 0.4529 mm, RMS 2.4973 mm, SD 2.4559 mm, and signed extrema $-16.1946/+16.2055$ mm. Two fingertip annotations are $+3.7031/+3.5113$ mm. Its ± 0.1 mm band contains 14.3326% of compared samples; this percentage describes that chosen surface tolerance. Signed averaging permits cancellation. The report's reference-model revision and complete registration project are unspecified, so those values describe its own geometric comparison. The current asset audit supplies additional, separately defined evidence.

The scene's original joint definitions match the registered older official Piper chain. For ten meshes, STL-to-OBJ nearest-vertex conversion error is at most 8.41×10^{-9} m. A 1,000-pose comparison of the small first-six-joint-origin differences yields at most 0.05874 mm at a point with the common 135.8 mm extension. These numerical checks locate the millimeter-scale discrepancy discussion at shape/version, assembly, pose, and observation layers. The newer official gripper chain includes a 4.5 mm flange term and a 138.0 mm finger-origin term, versus the older 135.8 mm combined definition; the 6.7 mm specification difference is a version comparison with its own meaning.

Registration and local identification. The analysis first fits a base transform, joint pose, and gripper opening under the existing link dimensions using robust, multi-start nearest-surface/point-to-plane fitting. It assigns visible surfaces to nearby links, then independently rigidly aligns selected link regions to diagnose local shape. The fitted joint angles are estimates. Local alignment supports a shape diagnostic; actual encoder angles and installation datums are required to identify full-arm kinematics and TCP uniquely.

In enabled pose 0.5, selected-region RMS before/after local alignment is 0.549/0.443 mm at the base, 0.555/0.219 at link2, 0.465/0.248 at link3, and 4.253/3.395 at the old gripper base. The persistent gripper-base residual motivates a version-specific geometric comparison. Crossbar widths extracted from three scans are approximately 165.395, 165.472, and 165.392 mm, compared with 145.0 mm in the older asset and about 164.083 mm in the newer official geometry. Width is the 0.1–99.9 percentile span of visible near-surface points in the fitted crossbar-height region, rather than an independent caliper measurement.

Two enabled poses yield four finger installation-origin estimates: 137.598, 137.502, 137.477, and 137.499 mm. The local calibrated asset uses 137.5 mm, retaining finger shape, axes, limits, and the

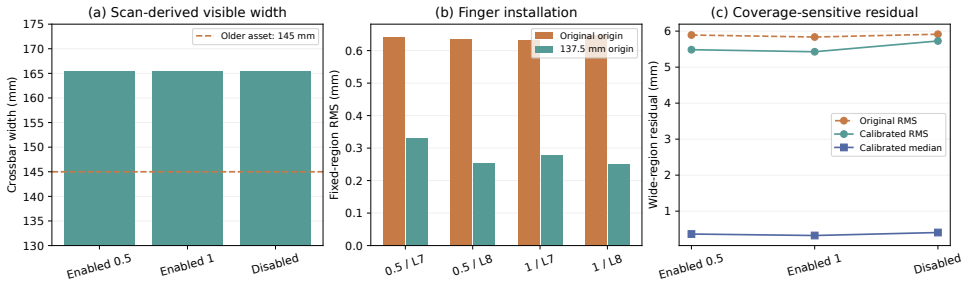

Figure S1: Asset-layer diagnosis from the supplied Piper scans and local calibration records. Left: crossbar-width estimates across three scans versus the older URDF geometry. Middle: fixed-region fingertip RMS before/after a 1.7 mm axial installation-origin adjustment. Right: wide-region gripper-base RMS and median residual retain their distinct meanings. Local visible-region improvement and full-region residual are reported together.

first six joints. Its gripper base uses the closer official template with a limited outer-crossbar width correction to about 165.395 mm; the central motor region is preserved. Visual and collision versions are updated together, and finger joint anchors and initial body transforms remain consistent. The local model retains its stated convex collision approximation.

Table S3: Fixed-region finger comparison. The two enabled poses and both fingers share the same test points before/after the installation-origin change. Values describe visible selected surface regions.

Pose	Link	Points	RMS before/after (mm)	P95 before (mm)	P95 after (mm)
Enabled 0.5	7	5678	0.642 / 0.332	1.357	0.784
Enabled 0.5	8	5794	0.635 / 0.255	1.339	0.475
Enabled 1	7	7331	0.632 / 0.278	1.525	0.638
Enabled 1	8	7241	0.649 / 0.251	1.659	0.520

The selected finger regions use proximity and local-coordinate masks with interleaved check points. Identification and evaluation share scan data; the second enabled pose is a cross-pose consistency check. In broader gripper-base regions, old/calibrated medians are 1.607/0.364, 1.322/0.323, and 1.852/0.406 mm; corresponding RMS values are 5.892/5.484, 5.836/5.427, and 5.914/5.726 mm. Roughly 28–30% of wide-region points lie outside the new template's 1.5 mm neighborhood, encompassing attachments, adjacent surfaces, and unresolved residuals. Reporting both median and RMS keeps this coverage visible.

Physical relevance and attribution scope. A separate calibrated-asset test executes 1,680 steps at 240 Hz (7 s): rest, light probe contact, small arm motion, gripper closing, reopening, and return. Collision queries hit the base and both fingers, 213 callback records capture probe contact, and terminal arm target error is about 0.000450 rad. The callback count is a sampling count. This verifies basic simulation control/collision operation for that asset. Mass, inertia, friction, the temporary camera-load estimate, and TCP retain their original status. The benchmark's archived policy results retain their original asset identity; the local calibration is an asset-validation study with its own records.

The resulting asset conclusion is specific: gripper-base version and local finger installation origin have consistent geometric evidence, while complete kinematic calibration, flange datum, scan repeatability, and true TCP remain separate identification requirements. Pose errors can amplify along a chain—for example, 0.5° over 0.5 m gives about 4.36 mm displacement—but this illustration is distinct from a measured joint error. The framework therefore traces a surface discrepancy to its geometric and observational layer before using it to explain contact or policy outcomes.

A.7 Visual index of real experiments and simulation assets

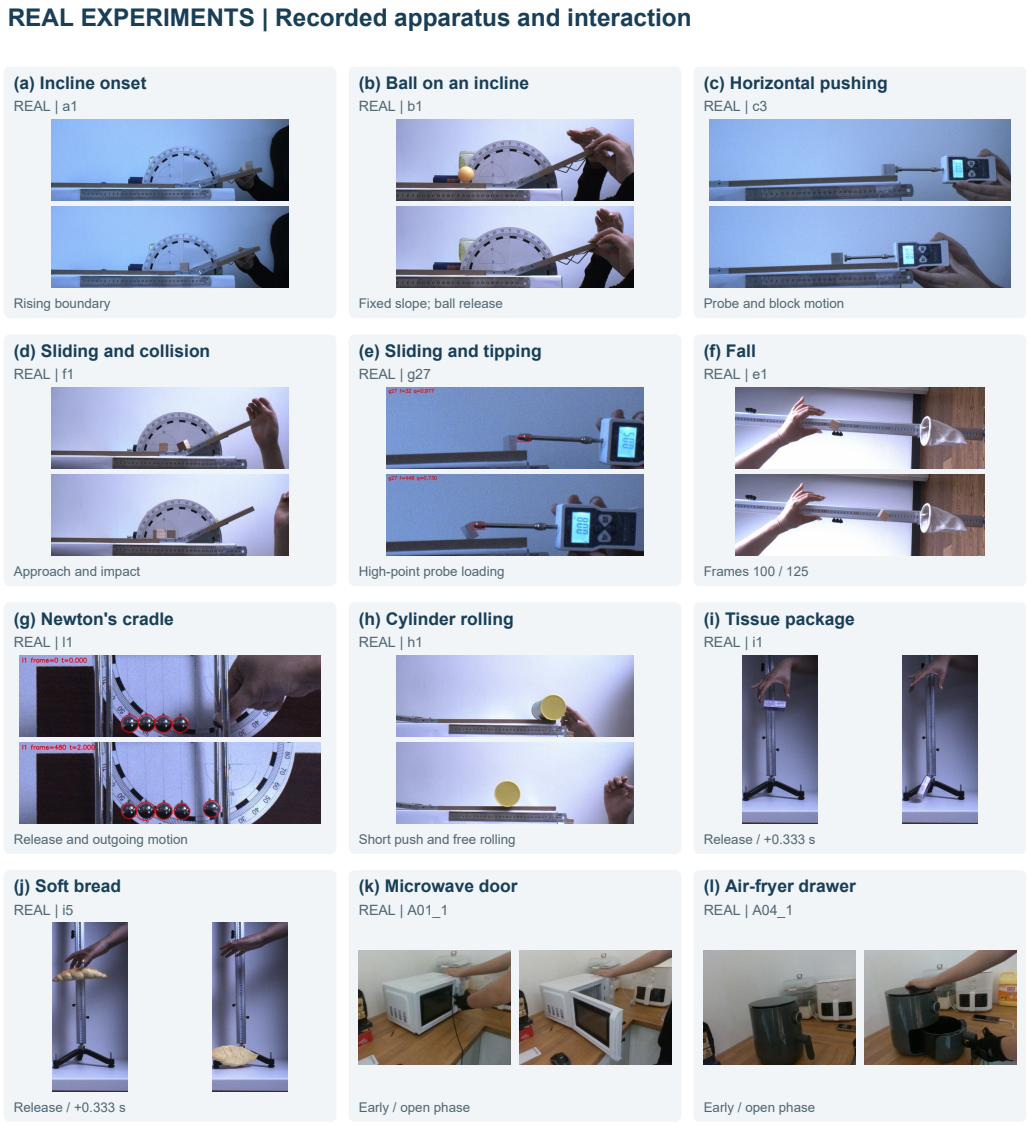

Figure S2: Consolidated real-experiment atlas. All panels are photographs from the registered recordings: (a) incline onset; (b) inclined-ball release; (c) horizontal pushing; (d) sliding and collision; (e) sliding–tipping competition; (f) fall; (g) Newton's cradle; (h) cylinder rolling; (i,j) tissue-package and bread release; and (k,l) microwave-door and air-fryer interaction. Two views are ordered top-to-bottom or left-to-right within each panel. Clip identifiers link to the corresponding records. Existing tracking overlays are retained in (e,g); (f) retains the recorded camera orientation. Insets have independent crops and scales. These views document apparatus and observable phases; quantitative evidence uses the registered trajectories, channels, and windows.

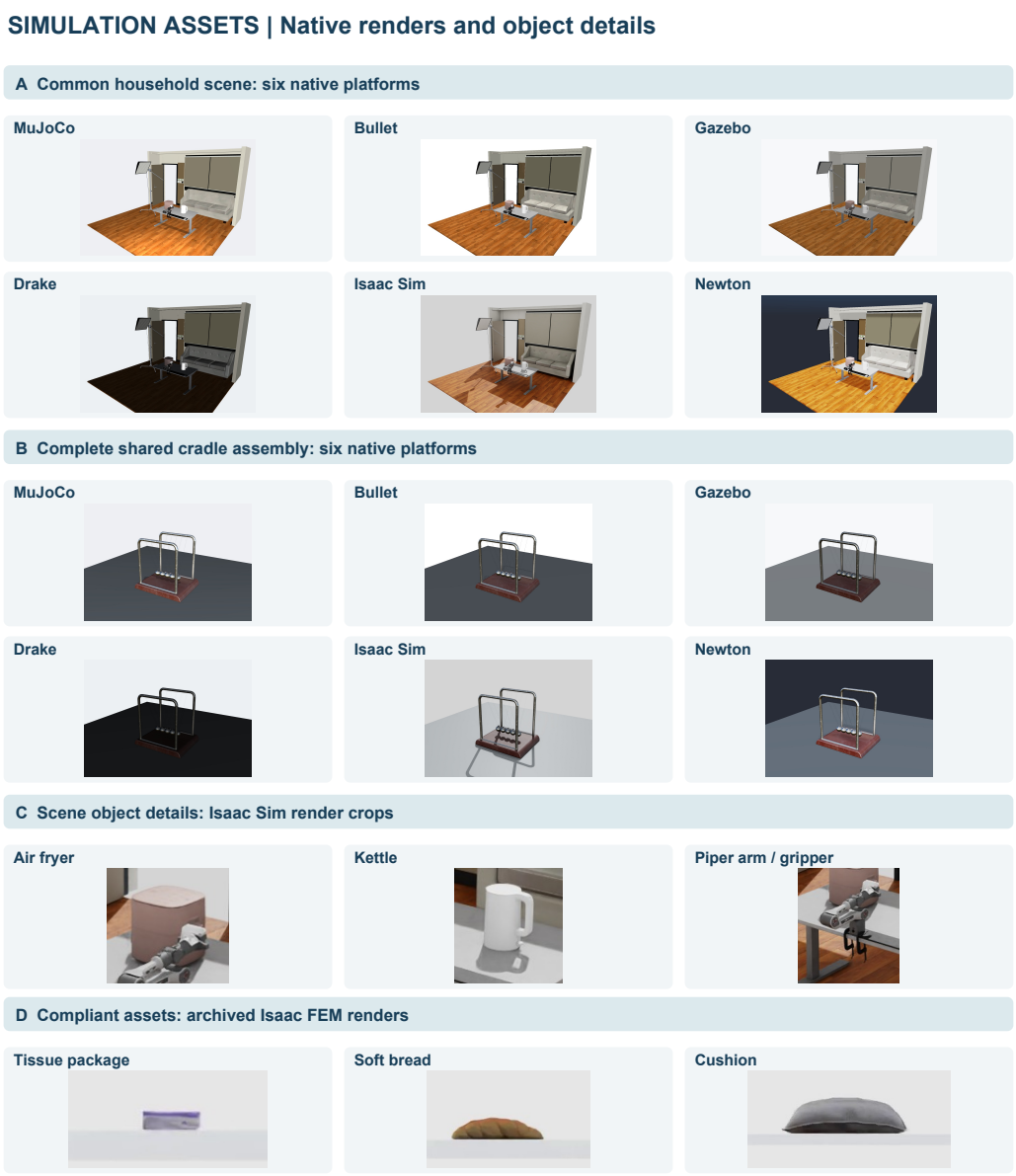

Figure S3: Consolidated simulation-asset atlas. Groups A and B show the common household scene and complete shared cradle assembly in six native renderers, respectively. Group C enlarges the air fryer, kettle, and Piper arm from the same Isaac Sim scene view. Group D shows archived native FEM renders of the tissue package, bread, and cushion. Platform labels, object identities, and independent display scales distinguish the views. The figures document visual assembly and asset representation; collision/constraint models and effective settings follow the registered protocols. Soft-body elasticity retains the stated material priors. Cradle mechanism controls retain the ideal-hinge configuration identified in Appendix B.3.

B Supplementary physical results and mechanism evidence

B.1 Aggregate support, onset, and free motion

The real onset, inclined-ball, and fall configurations are shown in Figure S2(a,b,f). The controlled

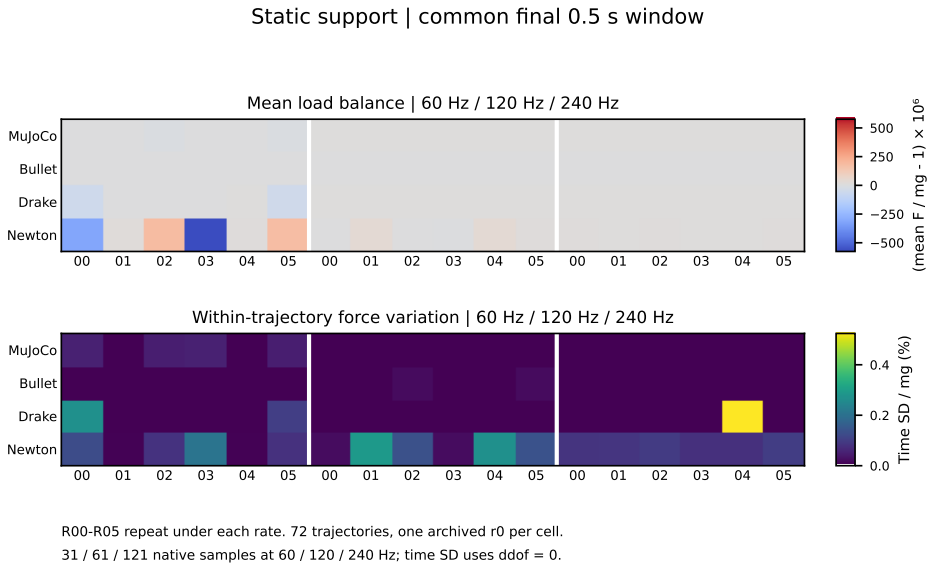

Figure S4: Common-window support analysis. Native channels preserve the distinction between summed support force and sensor magnitude. Mean equilibrium and within-window variation are separate observables; the recorded steady means support the main support score.

support archive has 324 registered observations, 54 per platform. In the common last-0.5 s subset, four platforms supply three-support force sums and Isaac supplies sensor magnitude. Gazebo's retained summary supports mean comparison while its cached continuous support array has a separate availability status. A representative time SD is about 0.52% of weight. This result explains how all mean scores can be 100 while fluctuation requires channel-specific analysis.

The a1 incline reference places real ramp-relative motion onset at 2.23–2.27 s. With 220 mm candidate ramp length, friction 0.30, and 240 Hz, five platforms confirm 4 mm displacement at 2.55–2.62 s; Drake's 1.35 s event accompanies about 87° rotation. The threshold, geometry, and mode jointly define the event. The base protocol's ramp rate and persistence contribute 0.2° confirmation delay. Horizontal force ramps show larger confirmed force under faster loading in 34 of 36 fixed-platform/friction/frequency pairs, with paired differences from -0.04 to 0.35 N.

At 240 Hz, e6/e9/e11 retain image residuals of 34.69/24.68/54.76 px over their respective evaluation windows. In a fixed later held-out window, increasing initial-state fit points from eight to sixteen changes residuals from 39.89/27.02/62.99 to 17.53/18.14/27.54 px. This is a same-window reconstruction control; time refinement provides an independent integration control. These paired questions retain separate denominator and window definitions.

B.2 Pushing, contact pairs, and rotational continuation

Figure S2(c–e) identifies the probe, contact geometry, and elongated-block arrangement used in the corresponding physical recordings. The g2/g27 contrast separates probe and floor friction. Upward probe tangential force changes $N \simeq mg - F_z$, modifying the floor friction budget; lateral contact changes force moment and mode. A representative 144-condition set crosses probe friction 0.10/0.30, height ± 2 mm, width $\pm 20\%$, and stretch 0.5/0.8/1.0 while retaining its specified units and source windows. Motion and contact responses are interpreted together with these registered factor combinations.

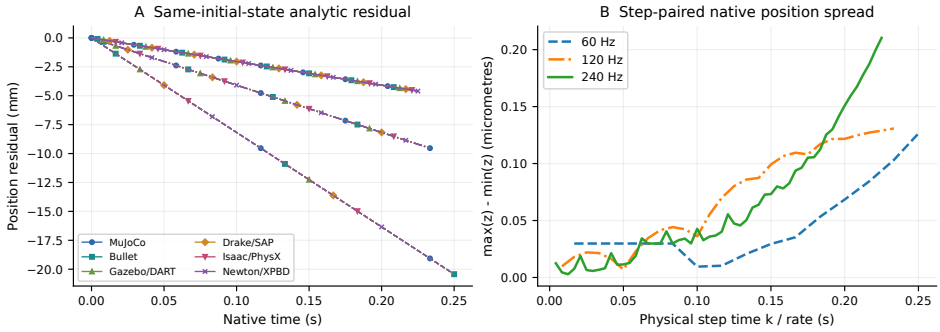

Figure S5: Fall residual and between-platform spread. Panel A separates physical frequencies by line style and platforms by color/markers. Panel B shows the matched-step maximum-minus-minimum height. Millimeter-scale common analytic residual and micrometer-scale platform spread use separate axes, explaining nearly coincident raw trajectories.

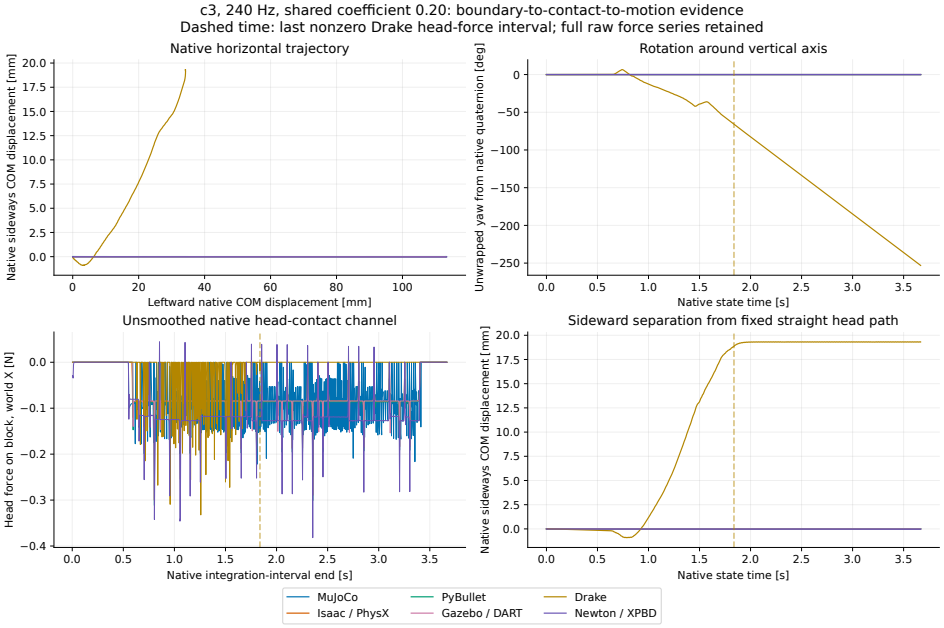

Figure S6: Drake c3 point-contact pushing. Boundary tracking, block lateral/yaw response, contact signal, and axial displacement share native time. Rotation/lateral motion precede probe separation at 1.84 s, followed by axial plateau. The sequence localizes the response to contact-induced motion under qualified input.

For a selected thin-body state near 92.39 rad/s, the registered 60 Hz continuation terminates with a nonfinite outcome, with its last finite sample at about 0.233 s; 120/240 Hz complete 0.25 s with energy ratios about 383.42/3.77. Equal-trace isotropic inertia retains bounded native states in the paired diagnostic. This combination identifies a sensitivity of the tested anisotropic rotational continuation and SolverXPBD configuration. The causal claim is conditioned on that saved state, inertia, and update path.

B.3 Cradle references, pair controls, and dissipation

Figure S2(g) shows the physical cradle and tracked balls.

The main micro-gap study uses ideal planar hinges with $D = 20.099795$ mm, $L = 128.989013$ mm, $m = 33.2$ g, $I = 2m(D/2)^2/5$, and gravity 9.80665 m/s^2. The separate thin-rod matrix uses its

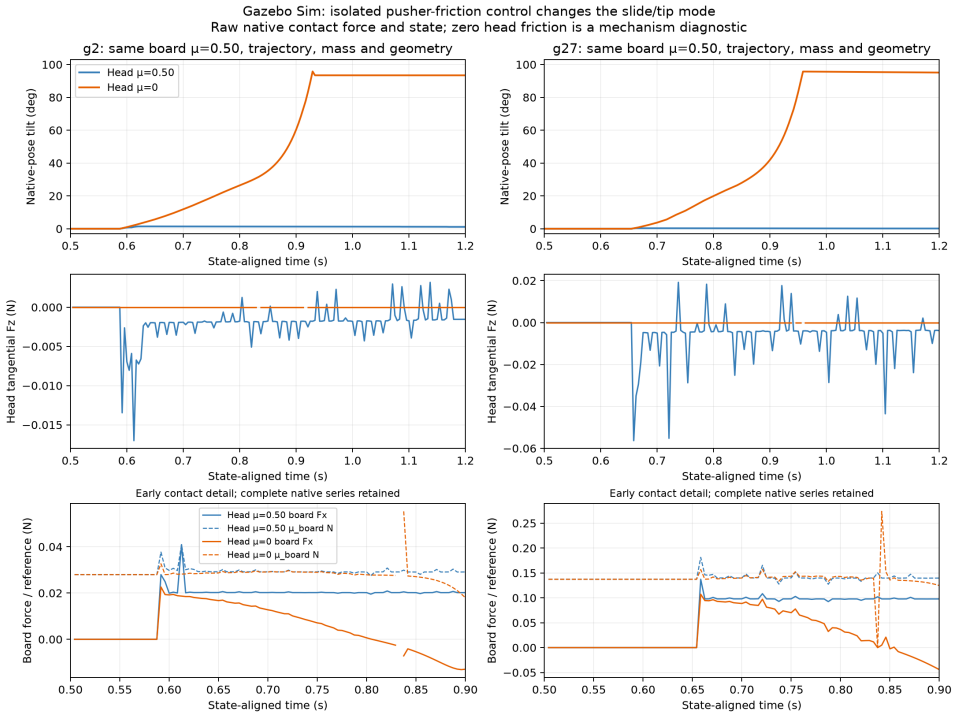

Figure S7: Gazebo contact-pair isolation with floor friction fixed at 0.50. Probe friction changes tangential load, body yaw, and contact mode in g2/g27. The actual boundary path and native body/contact records jointly define the controlled comparison.

Table S4: Six independent real releases. Angles are reconstructed from imagery; peak displacement is normalized by ball diameter D. Each row represents one physical release.

Condition	Released balls	Angles (degrees)	First peak (D)	Period (s)
l1	1	12.0	1.192	0.733
l2	1	28.6	3.125	0.733
l3	1	51.0	5.125	0.750
l4	2	13.3/13.3	1.442	0.717
l5	2	28.0/28.2	3.208	0.733
l6	2	48.9/50.4	5.025	0.750

registered 132.6519 mm effective geometry and 960 Hz protocol. For the ideal-hinge case, the first outward event is the first native sample satisfying $(x_1 - p_1)/D < -0.18$. Over $W = [t_o, t_o + 0.35\,\mathrm{s}]$, define

$$a_i = \max_{t \in W} \max(0, -[x_i(t) - p_i]/D), \qquad R_{21} = a_2/a_1.$$

The two maxima can occur at different times. R_{21} describes relative neighboring-ball excursion. The complete-asset views in the main text read the original `newtons_cradle_final.usda`: 18 rigid bodies, 20 revolute joints, and two fixed joints. Source meshes, texture coordinates, base, supports, ball eyelets, and ten cylindrical rods retain their original geometry. Analytic cylinders are tessellated from their source radii, lengths, and transforms for native display. Each of the six renderers receives the same rest assembly and camera; these views establish visual geometry correspondence. The quantitative five-planar-hinge model has its own identity and records. Its 60° release and the real reconstructed 51° release are separately stated. Figure 5 plots native positions without temporal smoothing, with each trajectory aligned to its registered outgoing event; the contact panel preserves absolute native times. The video curve uses retained observations and the archived occlusion reconstruction in one black trace; its source masks and fixed peak statistics retain their original definitions.

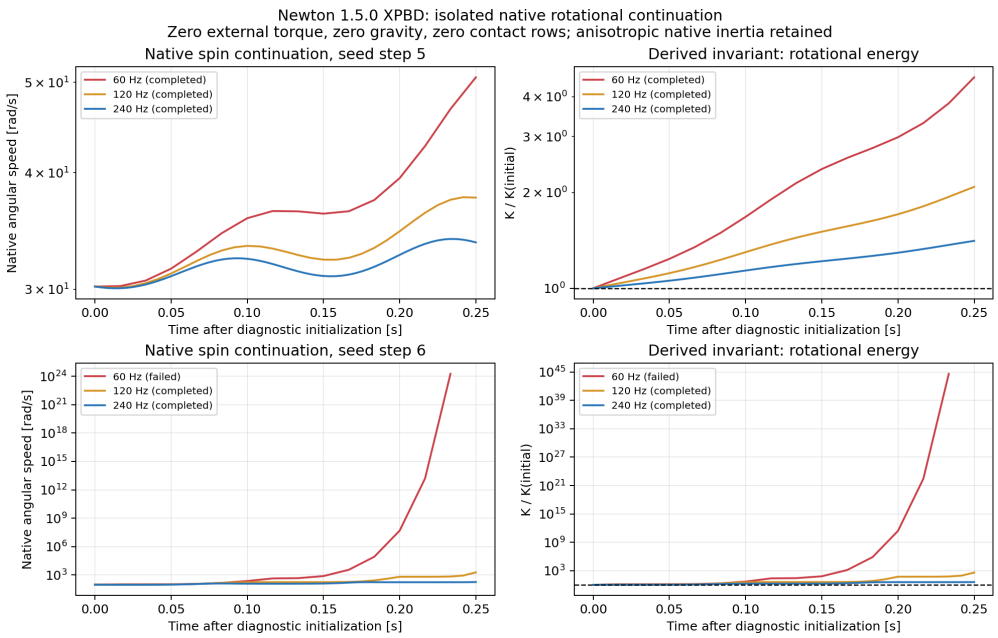

Figure S8: Native Newton isolated-rotation continuation from saved states. Gravity and contact are zero; left panels retain native angular state, right panels derive rotational kinetic energy. The reference energy is the continuation's initial value. Temporal refinement and inertia controls distinguish rotational propagation from the prior contact excitation.

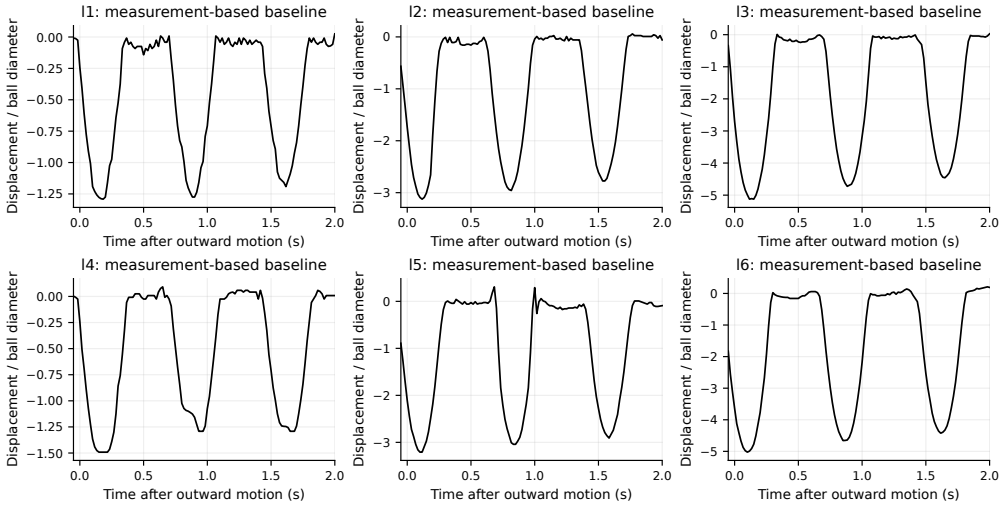

Figure S9: Measurement-based reconstructed reference baselines for six cradle releases, shown as continuous black curves. Short gaps use shape-preserving cubic interpolation; longer gaps use the measured cycle shape with endpoint-residual interpolation. The 600 retained observations and first-peak statistics remain unchanged. Displacement is diameter-normalized and time is relative to the first outgoing event.

Bullet's original post-step contact records supply positive normal force on all four adjacent pairs at 195.833 ms in the touching case: 4–5, 3–4, 2–3, and 1–2 carry 14.132, 10.652, 7.027, and 3.476 N. With a gap, the four activation times are 195.833, 200.000, 204.167, and 208.333 ms with 8.755, 8.778, 8.805, and 8.832 N. The plotted cutoff is 10^{-4} N. Timing resolution is one 4.167 ms phys-

Table S5: Ideal-hinge l3 at 240 Hz: touching / 0.1 mm gap. Each native configuration is one deterministic run. Potential-energy maxima use the full 22 s record and the 1.05 screening threshold.

Platform	First peak (D)	R_{21}	Maximum U/E_0
MuJoCo	3.905 / 4.824	0.657 / 0.645	2.797 / 1.477
Bullet	2.459 / 5.472	0.995 / 0.040	1.109 / 1.000
Gazebo	2.382 / 4.667	1.000 / 0.066	1.000 / 1.000
Drake	3.593 / 3.436	0.359 / 0.376	1.000 / 1.000
Isaac	2.379 / 5.043	1.000 / 0.045	1.000 / 2.756
Newton	1.355 / 0.880	0.750 / 0.943	1.000 / 1.000

ical step, with substep impulse details outside that observation scale. The repeated interface and preserved timestamps make contact sequencing directly comparable within Bullet.

For the isolated incident/output pair, an effective velocity ratio is $e_{\text{pair}} = (v_B^+ - v_A^+)/(v_A^- - v_B^-)$ using registered pre/post windows. Bullet velocities are $(0, -1.116225)$ m/s before and $(-1.087794, -0.017534)$ m/s after, giving 0.958821. Drake's registered SAP point stiffness 10^6 N/m and relaxation 0.001 s give 0.029421. This is a windowed dynamic measurement under the corresponding hinge/gravity configuration. Coupled five-ball outcomes then add contact topology and simultaneous-constraint effects.

The real l3 envelope is approximately 2.325 D at 10 s and 0.908 D at 20 s. At 240 Hz, Bullet's gap case gives 2.094/0.920 D, versus 3.200/2.074 D when touching. Full-window energy screens retain separate warnings: the ideal-hinge maxima exceed 1.05 in both MuJoCo cases, touching Bullet, and gap Isaac. First-peak agreement and passive-energy behavior therefore remain separate conclusions. Native hinge-length error reaches about 0.522 mm for the tested Newton case, versus at most 0.00166 mm for Isaac; initial pose readback is below 15 nm across the twelve selected records.

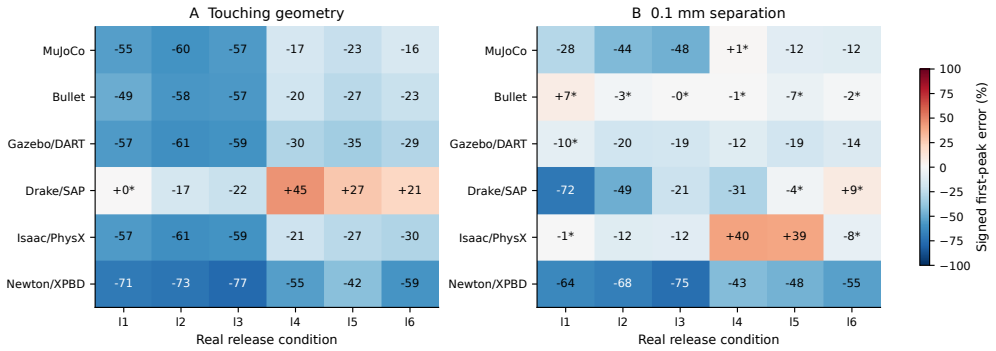

Figure S10: Separate 960 Hz thin-rod release–gap response matrix. The common signed-relative-peak scale enables within-model paired comparison across six real releases. This matrix supplies cross-condition corroboration of geometric sensitivity; its geometry and frequency are registered separately from the main 240 Hz ideal-hinge case.

In that matrix, Bullet's six first peaks move from 0/6 to 6/6 within the 10% screening band. The l3 peak changes from 2.20 to 5.11 D; Gazebo changes from 2.08 to 4.17 and Isaac from 2.08 to 4.53. Drake's mean error increases in the paired gap set, illustrating configuration-dependent direction. The real l3 reference is 5.13 D.

B.4 Native numerical controls

The registered Drake 12-face, 1 MPa compliant representation loses terminal support at 60 Hz; a 52-face counterpart supports approximately 1.29 N, and the softer 0.1 MPa control also retains support. MuJoCo contact-time constants from 0.04 to 0.10 s change effective bounce ratio from about 0.82 to 0.70 while peak penetration changes from 7.33 to 16.68 mm. Within the same tested condition,

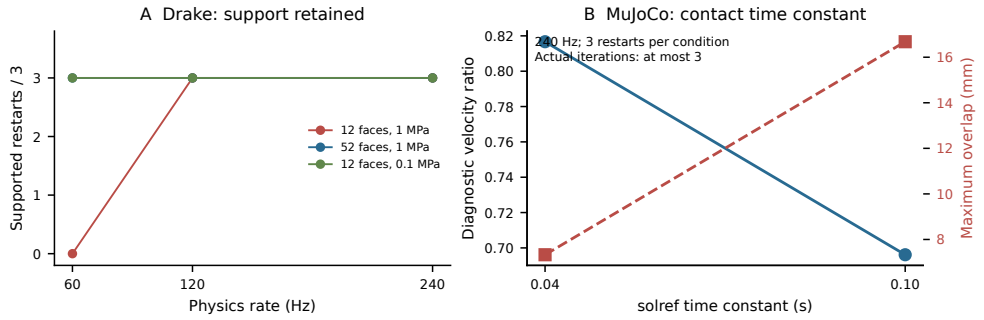

Figure S11: Representation and numerical-budget controls. Left: Drake support retained under mesh/contact representation and time changes. Right: MuJoCo restitution-window ratio and penetration under contact time-constant changes. Matched geometry and input define each local comparison.

budgets 10/50/200 can yield the same actual stopping count, which illustrates the distinction between requested and used iterations.

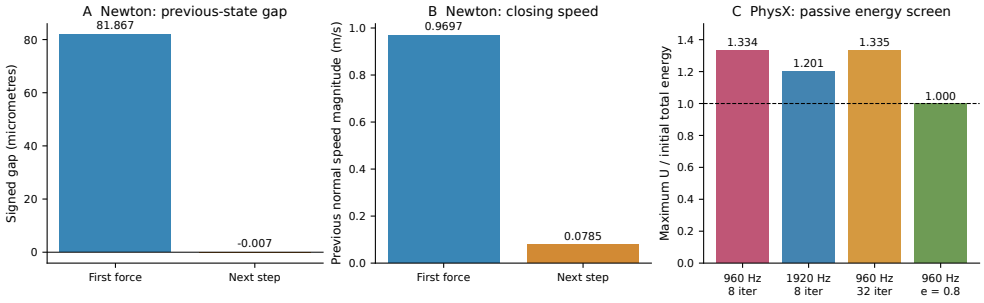

Figure S12: Local native restitution and energy diagnostics. Newton panels retain the previous-state gap and closing speed that govern activation; PhysX panels compare full-window energy behavior under registered numerical/restitution settings. These controls support branch- and configuration-specific interpretations.

The recorded Newton example moves from an approximately 1.87 μm separated prior state with 0.9697 m/s closing speed to near-touching after constraint correction with 0.07848 m/s closing speed. Changing contact margin from 10^{-5} m to zero shifts the sampled branch and recovery behavior. The paired evidence connects timing to native restitution evaluation. Selected PhysX thin-rod controls change maximum energy ratio from 1.3343 at 960 Hz to 1.2011 at 1920 Hz; lowering nominal restitution from 0.95 to 0.80 gives approximately 1.0000. Energy screens jointly reflect constraint/contact coupling and numerical configuration.

C Household compliance and tactile observations

C.1 Shared inputs and native deformable representations

Figure S2(i,j) shows the physical release observations; Figure S3, group D, separately shows native FEM assets under the stated model priors. The soft-body set includes bread-like packages, tissue packages, and milk-container proxies with nominal release heights 0/0.2/0.4/0.5 m, flat/vertical placements, mass inputs spanning approximately 0.020–0.587 kg, floor friction 0.2/0.5, and elastic-modulus priors 40/200 kPa. The common coarse surface and tetrahedral representations, density, and zero material damping are retained with each asset. Native methods comprise MuJoCo flex with implicit integration, Bullet Neo-Hookean tetrahedra, Drake corotated FEM with SAP contact, New-

ton XPBD for single bodies and VBD for stacks, and Isaac tetrahedral GPU FEM. Gazebo's installed tested DART path has a separate volumetric-soft-body interface coverage status.

Each single body advances for 1 s; each stack for 0.7 s. Before execution, records include mesh counts, mass, minimum initial height, surface pair, and material priors. During execution, available node positions, velocities, contact states, signed tetrahedral volume, and penetration are retained. End evaluation checks finite state, positive orientation, maximum penetration, settling/compression, contact count, and wall-clock cost. Full-dimensional material accuracy depends on measured constitutive parameters; the present comparison measures response to shared stated priors.

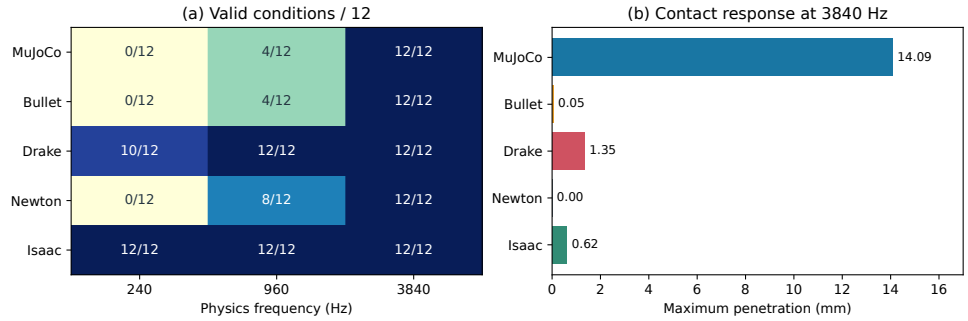

Figure S13: Soft-body numerical validity and contact response. Twelve exact conditions per frequency define each validity denominator. The penetration panel uses the specified high-rate condition set and native node coordinates. Positive-volume and finite-state tests qualify the subsequent contact interpretation.

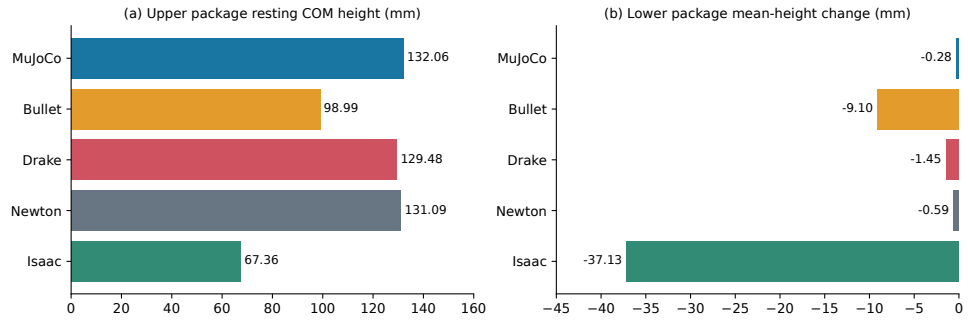

Figure S14: Two-package stack comparison. The left panel records upper-package resting center-of-mass height; the right records the lower package's subsequent mean-height change. Native deformation and inter-package support jointly contribute to these task-conditioned responses.

For the registered stack, upper-package heights are 132.06/98.99/129.48/131.09/67.36 mm for MuJoCo/Bullet/Drake/Newton/Isaac; lower-package changes are $-0.28/-9.10/-1.45/-0.59/-37.13$ mm. A separate coarse-particle Newton XPBD stack configuration places the falling package's terminal center of mass near 14.46 mm, whereas the registered VBD configuration supports surface contact. At fixed 3840 Hz, changing MuJoCo contact time from 10 to 1 ms reduces three selected maximum penetrations from 13.980/14.955/14.426 to 9.584/10.177/9.905 mm. These outputs distinguish spatial/constitutive response from temporal/contact controls.

At 3840 Hz, maximum single-body penetration is 14.09, 0.05, 1.35, approximately zero, and 0.62 mm for MuJoCo, Bullet, Drake, Newton, and Isaac. Native corotated, Neo-Hookean, XPBD, VBD, and FEM paths retain their own constitutive meanings. These numerical details complement the phase-level comparison in the main text.

C.2 Real tactile collection and phase-conditioned quantities

The real collection contains 32 conditions with ten operations each (320 operations): six gripper-opening tasks, three bowls, two trays, six noodle-like objects, three dishes, four bread conditions, one box, three tissue conditions, three bread clips, and one bottle-cap condition according to the recorded asset grouping. The five-finger interface retains six-dimensional node messages where available and raw per-finger intensity. Derived quantities include peak, integral, P95, dispersion, duration, opening/closing hysteresis, and branch-specific slope. Original raw-unit channels and calibrated physical channels retain different unit tags.

For a calibrated sensor node with raw vector $\mathbf{r}_i$, offset $\mathbf{b}_i$, gain map A_i, frame rotation R_i, and position $\mathbf{p}_i$, the physical mapping would be

$$\mathbf{f}_i^O = \sigma_i R_i A_i(\mathbf{r}_i - \mathbf{b}_i), \quad \mathbf{F}^O = \sum_i \mathbf{f}_i^O, \quad \boldsymbol{\tau}^O = \sum_i (\mathbf{p}_i^O - \mathbf{p}_{\text{ref}}^O) \times \mathbf{f}_i^O.$$

The current uncalibrated real intensity is retained in unit U for stage-conditioned comparisons. Its integral has units U s. Node calibration, orientation, and moment arms are required for a physical force/torque interpretation.

Reception timestamps pair tactile data and RGB-D by a 30 ms nearest-neighbor rule. Tactile data are approximately 100 Hz, RGB-D approximately 30 Hz, and geometric opening about 5 Hz. Geometry gaps above 60 ms carry continuity flags. A typical 0.4 s opening contains about two geometric observations; broad phase comparisons follow that resolution. Opening bands and quintile/time-normalized summaries retain branch identity. Each trial is summarized first; between-trial SD, IQR, or 10,000-resample intervals use trials as the sampling unit. Geometry continuity qualifies 21/60 trials in the separately registered six-condition geometric subset.

D Appliance articulation and independent dynamic validation

D.1 Roles of asset, motion input, force, and tactile reference

Figure S2(k,l) shows the corresponding real opening actions and hand contact. The appliance study uses engineering articulation models, measured opening trajectories, theoretical torque inputs, and real hand observations for distinct purposes. Shared URDF-converted geometry establishes model and input consistency. Replaying a measured opening trajectory tests controlled geometric response. An isolated eccentric hinge with prescribed inertia and torque tests dynamic response against an independent recurrence. Real fingertip intensity describes hand support and operation stage. This division separates trajectory tracking, physical-effort prediction, and observation semantics.

Table S6: Shared appliance mechanisms. Masses are engineering model inputs for the moving part. Native angular storage uses radians.

Asset	Joint	Mass	Range and modeled part
White microwave A01	Hinge	1.0 kg	Door, 0–110°, common door inertia
Gold microwave A02	Hinge	1.2 kg	Door, 0–100°, stated mass/inertia
Air fryer A04	Slider	0.6 kg	Drawer, 0–160 mm, observed local-displacement proxy
Breakfast appliance A05	Hinge	0.8 kg	Lid, 0–90°, standard independent torque input

The articulated microwave/air-fryer validation assets in this table and the scene-showcase appliance assets have separate identities and purposes. Native adapters preserve mass and inertial tensors, joint axes, local anchors, and transformed centers of mass. Mesh and inertia transforms use the registered SI, Z-up convention. For example, the white microwave door's center of mass is $(-0.177149, 0.009899, 0.000162)$ m in its joint frame, with vertical-axis inertia 0.04165696 kg m^2. These values are engineering specifications rather than real inertial measurements.

D.2 Control, timestamps, and interface readback

Gravity is $(0, 0, -9.81)$ m/s^2. Adapters compute the same clipped PD input from current joint state and advance one native step:

$$\tau_k = \text{clip}\{k_p(q_k^* - q_k) - k_d\dot{q}_k, -\tau_{\max}, \tau_{\max}\}.$$

Hinges use $k_p = 15$ N m/rad, $k_d = 3$ N m s/rad, and 8 N m limit; sliders use 1000 N/m, 40 N s/m, and 120 N. Limits define the common control protocol. State, input, target, and native force/constraint outputs share step intervals at 60/120/240 Hz. MuJoCo/Bullet/Newton request 100 iterations; Gazebo/DART and Drake retain their documented solver paths; Isaac uses TGS budgets of 8/4. Newton's native XPBD linear/angular relaxation defaults are 0.7/0.4 in the registered comparison.

MuJoCo joint position/velocity and generalized applied/constraint/bias forces retain their separate fields. Bullet retains torque-control input, state, and reaction wrench. Gazebo retains pre-update input and post-update state/wrench with phase labels. Drake retains generalized input and discrete contact/constraint results. Isaac retains articulation state, measured joint forces, and active effort. Newton retains body/joint states and its original reaction/correction meanings. Derived joint projections use the corresponding axis, wrench origin, and common interval.

The real collection supplies 13 geometric inputs: five A01, six A02, and two A04. Five/four/two respectively satisfy continuous-input criteria for grouped RMSE. All 13 enter the native reproduction queue, producing $13 \times 6 \times 3 = 234$ input-following runs. The standard-input queue contributes $4 \times 6 \times 3 = 72$ runs. An additional 102 same-window force comparisons produce 408 total runs in these registered groups. There are 51,764 posture states, 1,051,764 control/contact records, and 1,401,408 finite-state samples. Each trial retains its input continuity window and geometry semantics.

At 240 Hz, mean tracking RMSE is approximately 4.632–4.640° for A01, 4.445–4.456° for A02, and 1.873 mm for the A04 local-geometry proxy. Force comparison uses recorded native acceleration when exposed, or phase-labeled velocity differences when derived. Matching motion under the common controller can require different generalized input. MuJoCo theoretical/input effort consistency is about 0.012300–0.012311 N m in the selected white-door control, while the corresponding observed platform effort differences remain individually recorded.

D.3 Independent eccentric-hinge test

With a vertical hinge axis and no active stop contact, the angular reference is

$$J = \mathbf{a}^\top I_C \mathbf{a} + m\|\mathbf{r}_C \times \mathbf{a}\|^2, \quad \omega_{k+1} = \omega_k + h\tau_k/J, \quad q_{k+1} = q_k + h\omega_{k+1}. \tag{6}$$

Here $\mathbf{a}$ is the unit joint axis, I_C is center-of-mass inertia, and $\mathbf{r}_C$ is the joint-to-COM vector. The comparison uses the same semi-implicit recurrence as its declared discrete reference. The model starts at $q = 0$, applies 0.01 N m for 0.5 s, then coasts for 0.5 s. At 240 Hz, the reference endpoint is about 0.090271 rad, with continuous-time limit about 0.090021 rad. Model mass/inertia and input are fixed during the numerical-parameter sweeps.

Table S7: Newton native hinge controls, one deterministic run per exact configuration. Relative endpoint error uses the matching-step analytic recurrence.

Factor	Values	Endpoint deviation (%)
Iteration budget	10/50/100/200/400	47.604, 47.711, 47.711, 47.713, 47.716
Physics frequency	60/120/240/480 Hz	44.826, 47.139, 47.711, 47.833
Centered COM control	Zero offset	about $+0.771$
Angular relaxation	0.7 with linear 0.7	about -0.326
Matched relaxation	0.4/0.4; 1.0/1.0	about -0.185; -1.442
Inertia updated from geometry	Registered alternative	about -46.433
Velocity/position re-projection	Registered native control	about -48.079

The paired relaxation response, matched inertia/input, and independent recurrence jointly support a configuration-specific mechanism. At the same point in the input pipeline, five other tested backends

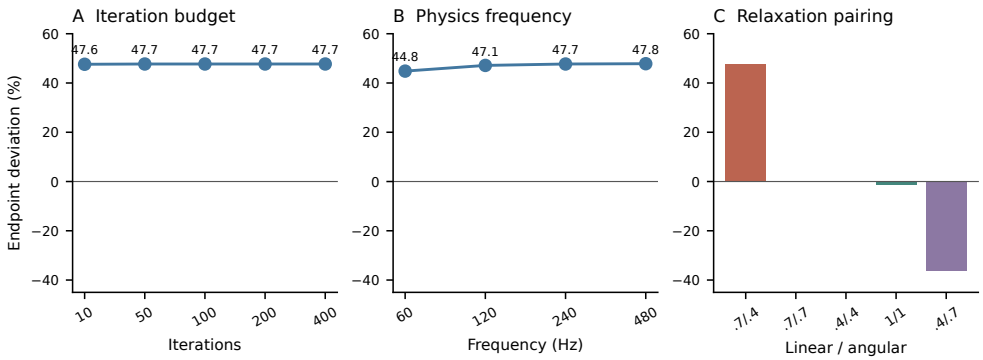

Figure S15: Iteration, physical-frequency, and relaxation controls for the isolated hinge. Distinct responses to these factors localize the tested sensitivity to eccentric articulation and the native relaxation configuration. Iteration budget and temporal resolution are separate axes.

produce endpoint error below 0.001% under their registered hinge models. The result describes this articulated configuration and its control path. A separate breakfast-appliance standard opening supplies a geometric replication on another asset.

D.4 Tactile phases and evidence scope

The tactile appliance subset has 20,980 node messages and 125,880 node records across 13 inputs. Both white and gold microwave operations show opening/closing responses and lower holding-phase means. Selected white-door opening-phase means are 6.50–11.76 U, closing-phase means 2.68–4.50 U, and holding-phase means 0.15–1.38 U. Air-fryer A04 uses a local displacement proxy associated with pulling and support phases. Its two post-closing means are about 11.159 and 0.810 U; the respective holding means are 2.432 and 11.800 U. Geometric definition and sensor phase determine which comparison is meaningful.

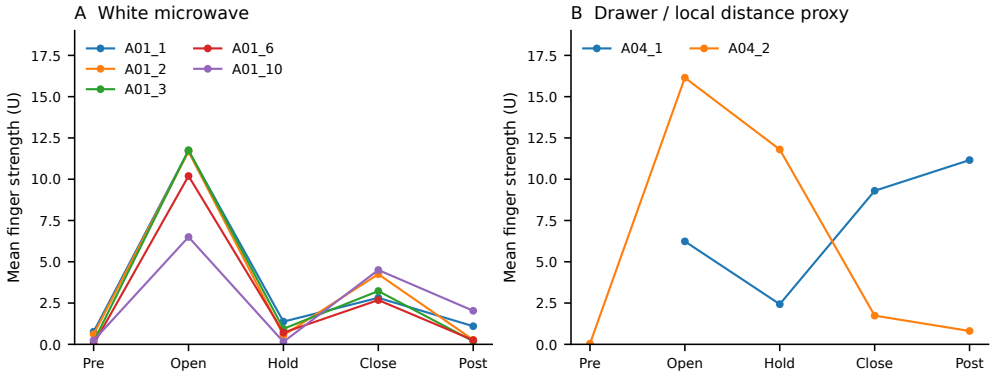

Figure S16: Trial-wise real tactile phase observations. Each trace is one operation in raw intensity units U. Phase means describe support and opening/closing activity. Geometry and native generalized torque define different physical quantities, with their relation mediated by hand contact points and calibration.

The final evidence has three layers: common-input readback supports comparability; the eccentric-hinge reference and parameter controls support a native dynamic mechanism; and real hand signals independently support the staged interaction observation. Contact locations, complete hand wrenches, and calibrated force mappings define the scope of a direct hand-force-to-joint-torque comparison.

Table S8: Minimum record groups and their explanatory roles. Field names retain native API definitions within each adapter.

Record group	Retained fields	Explanatory role
Condition	Task/platform/version, geometry, inertia, material pairs, h, n_s, K, readbacks	Identifies common and native-effective inputs
State	Time/phase, pose, velocities, joint state, node positions	Locates onset, drift, rotation, deformation
Contact	Body pair, point, normal, gap, native response, coordinate origin	Connects contact establishment to force/impulse
Control	Target, raw and clipped command, actual applied value, rate	Separates action generation from execution
Constraint/solver	Native residual/correction/iteration fields where exposed	Tests numerical and articulation mechanisms
Sensor/transport	Raw values, source and reception time, validity/coverage	Aligns observation with physical phase
Outcome	Event times, completion, finite state, reason, usable window	Defines denominators and terminal conditions

E Channel dictionary and reproducible computation

E.1 Interval alignment and force aggregation

Each external step $[t_k, t_{k+1}]$ associates the input actually executed with that interval's native contacts and the resulting state. A contact query may expose pre-step geometry and interval-solved force, or a completed-step contact record; these meanings are preserved. Impulses accumulate over the physical interval before conversion to mean force. Forces rotate into a common frame and use a consistent sign relative to the evaluated body. Torques additionally require their moment reference. Solver, contact, and sensor aggregations keep their original spatial support.

When a complete interval force set is available, the main momentum residual checks consistency with measured velocity change. Partial channel coverage is identified as such. Raw force zero, unavailable channel, and unavailable contact frame have separate statuses. Newton correction/impulse-derived quantities, MuJoCo generalized constraint force, Bullet contact-point force, DART wrench, Drake contact force, and PhysX contact/sensor outputs enter the comparison through their own mapped definitions.

Real geometry uses source video time; asynchronous tactile/RGB-D uses the reception-time rule already described. Interpolation is qualified by continuity and sample spacing. Contact aggregation, body-state sampling, and event confirmation are computed on registered source intervals. High-frequency raw records remain distinct from stage means and presentation smoothing.

E.2 Evidence strength across tasks

The attribution hierarchy uses increasingly discriminating evidence: input readback qualifies the cause space; first-divergence timing suggests a stage; one-factor response tests a candidate; independent mechanics and native implementation constrain the mechanism. In pushing, qualified probe motion plus yaw/separation localizes contact-induced drift. In free fall, shared recurrence plus fixed-window reconstruction separates time integration and image-state estimation. In the cradle, common poses plus gap pairs, adjacent-contact timing, and two-ball response support topology-dependent transmission. In articulation, common inertia/torque plus a known recurrence and relaxation sweep support the native configuration effect. In policy execution, command/tracking/event observations distinguish approach-limited cohorts from trained policies that reach contact, holding, and release (Appendix F). These evidence types make each conclusion specific to its observed phase and tested conditions.

F Six-platform policy training and closed-loop execution

F.1 Comparison design and shared learning protocol

The learning comparison follows the benchmark's separation of shared conditions, platform-dependent responses, and process attribution. Six source-platform datasets train six PiPER $\pi_{0.5}$ LoRA policies for pushing, picking, and placing. Each trained policy is evaluated on its source platform. This design measures the combined effect of platform-conditioned demonstrations, learned actions, and native execution. The separately registered cross-source execution matrix isolates a different comparison axis and retains its own cohort identity below.

The balanced dataset contains 72 training and 24 validation episodes per platform: three tasks, four materials, and six/two scenarios per task–material combination. A demonstration scenario enters the balanced set when the same scenario passes on all six backends. All platforms therefore share scenario identities and task definitions, while each backend supplies its own native trajectory and observations. This selection establishes demonstrated task feasibility over the common scenario intersection. Training and validation have disjoint episode identities; the normalization statistics are fitted once on the combined six-platform training split. The closed-loop evaluation uses a held-out aluminum scenario with seed 990001; aluminum also appears in training, so the holdout is at the scenario level.

Within each source platform, the 4,000- and 20,000-update runs share the initial base weights, data manifest, normalization, validation frames, training module, and effective-channel objective. Each run uses training seed 42, batch size one, peak learning rate 2.5×10^{-5}, terminal rate 2.5×10^{-6}, and 100 warm-up updates. Task-balanced sampling assigns the registered weights 2.0 to closure/lift/transport/release, 1.0 to approach/retreat, and 0.75 to hold/settle. Validation uses 12 fixed frames per task, uniformly selected and evaluated every 200 updates with recorded noise settings. The comparison treats update budget and its associated training schedule as the training condition.

The architecture retains a 32-dimensional flow representation; the six arm channels and gripper channel define the supervised action coordinates. For predicted and target flow velocities v_i, u_i, the recorded objective is

$$\mathcal{L}_{\text{active}} = \mathrm{E}_{x,t,\epsilon}\left[\frac{\sum_{i=1}^{32} w_i (v_i - u_i)^2}{\sum_i w_i}\right], \qquad w_i = \begin{cases} 1 & i \le 7, \\ 0 & i > 7. \end{cases} \tag{7}$$

This normalized flow loss characterizes prediction on the retained validation frames. Physical accuracy is measured separately through native states, contact events, and task outcomes. Across source platforms, common sampling rules and normalization coexist with platform-specific demonstrated trajectories; their losses describe learning on those respective data distributions.

Inputs are head and wrist RGB, six measured joint angles, gripper width, and task text. Inference produces 16×7 absolute targets and executes the first eight at 15 Hz; native state is recorded at 240 Hz with four integration substeps (960 Hz). Common arm target-rate and gripper-width-rate limits are 1 rad/s and 0.025 m/s. The 25 mm cube has the registered aluminum mass of approximately 43 g, table-pair friction 0.4, and gripper-pair friction 0.8. Joint gains, effort limits, geometry identities, and initial scenario are retained; native solver settings retain their backend meanings. Robot–object–table contact is advanced by each backend's dynamics.

F.2 Training trends and task-conditioned completion

All six sources have complete finite records of 4,000 and 20,000 updates, with matching per-update CSV and structured logs. Table S9 and Figure S17 connect the common training protocol to the observed loss trends. Final validation loss decreases by 63.93–76.14% between the two budgets across the six source datasets. Both budgets start from the registered base weights; the plotted 4k run and the first 4k updates of the 20k run are separate optimization trajectories.

The paired evaluation scenario yields 6/18 completed tasks at 4k and 13/18 at 20k. All six pushing cases pass at both budgets. Picking changes from 0/6 to 4/6 and placing from 0/6 to 3/6. Thus, the observed increase concentrates in sustained grasping and multistage placement, while the easier pushing cases already satisfy the task criterion. These are observed counts for one scenario per

Table S9: Source-platform training and source-platform execution. Validation flow loss is shown multiplied by 10^3; completion counts cover three tasks per platform.

Platform	4k loss	20k loss	Reduction (%)	4k pass	20k pass
MuJoCo	3.882	1.114	71.31	1/3	2/3
Bullet	4.103	1.094	73.32	1/3	2/3
Gazebo	4.420	1.594	63.93	1/3	2/3
Drake	4.852	1.230	74.66	1/3	2/3
Isaac Sim	5.785	1.380	76.14	1/3	3/3
Newton	4.492	1.087	75.81	1/3	2/3

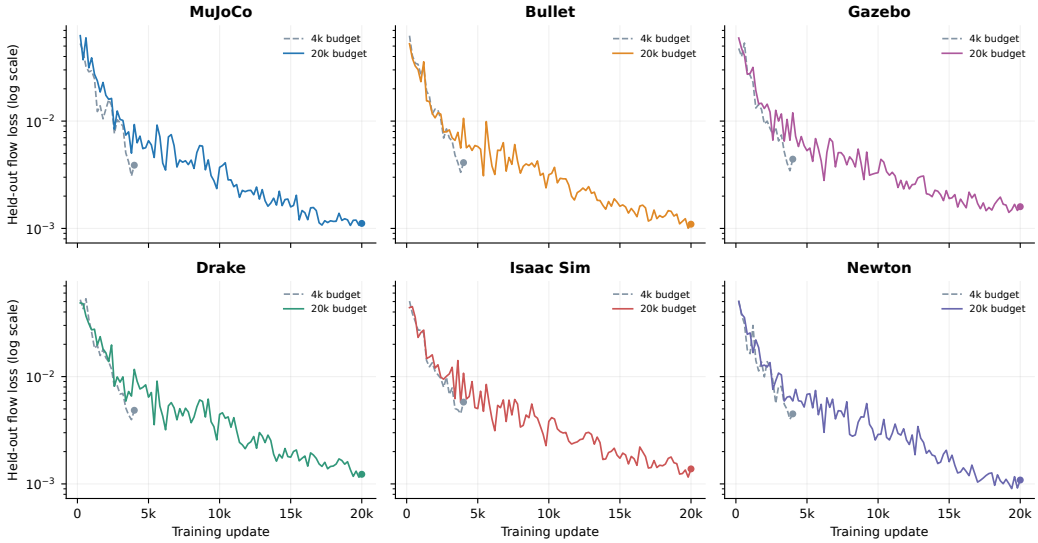

Figure S17: Validation flow-loss trajectories for the six platform-conditioned training datasets. Solid and dashed curves are the 20k- and 4k-budget runs; dots mark their endpoints. Every plotted sample is a recorded validation evaluation, with a common logarithmic vertical scale and no temporal averaging. Shared scales support trend comparison; source-specific trajectories determine the underlying learning distribution.

platform–task pair and one training seed. Each cell represents a complete episode rather than a repeated-trial probability estimate.

The training result and the task matrix provide complementary evidence. Loss improves on all six datasets, whereas complete grasping and placing retain distinct outcomes. The explanatory object is therefore the learned-policy–controller–backend pipeline at a specific interaction stage. The common protocol establishes comparable learning conditions; continuous execution traces identify which stage accounts for the remaining differences.

F.3 Inference records and stage-resolved platform differences

Every evaluated rollout retains model inputs, raw predicted chunks, applied targets, measured joints, object pose, native contact records, and inference timings. The action target is associated with the physical interval in which it is held. Two useful summaries are

$$e_q = \sqrt{\frac{1}{6N}\sum_{k=1}^{N}\sum_{j=1}^{6}(q_{k,j} - q_{k,j}^{\text{applied}})^2}, \qquad S_{3s} = \left\| \max_{t\in[T-3,T]} \mathbf{p}(t) - \min_{t\in[T-3,T]} \mathbf{p}(t) \right\|_2. \tag{8}$$

The maximum and minimum in S_{3s} are componentwise. Thus e_q describes arm execution relative to the command actually applied, while S_{3s} describes final object stability. Gripper target/measurement

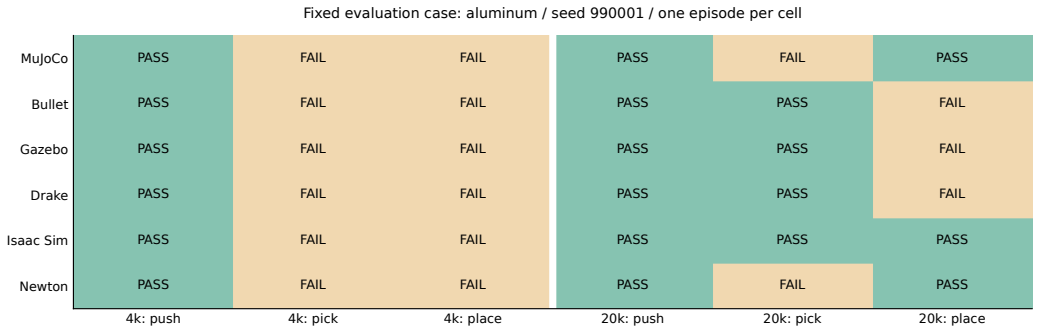

Figure S18: Task-by-platform completion at two training budgets. Each cell is one closed-loop rollout, using the same evaluation scenario identity across budgets. The matrix makes the common pushing outcome and task-specific grasp/placement differences explicit. It evaluates each source-trained policy on its own native backend.

differences remain separate because closing onto a finite-width object has a physical contact interpretation. Native force/impulse fields retain their interface semantics; the common outcome tests use object positions, measured finger joints, and explicitly derived shared-URDF geometry.

Grasping: common access to lifting, distinct holding responses. All six 20k policies finish with lift above the 40 mm target. Their final lifts span 59.31–70.28 mm. The complete-task difference lies in the final 3 s stability criterion: Bullet, Drake, Gazebo, and Isaac Sim have spans below 5 mm; MuJoCo and Newton have spans of 5.165 and 8.196 mm. The MuJoCo case lies only 0.165 mm above the common boundary. Reporting this continuous margin preserves the distinction between a near-threshold result and a larger physical excursion.

Table S10: Continuous pick metrics at 20k. Hold span uses the final 3 s of native object positions; arm RMSE compares measured joints with applied targets, in mrad.

Platform	Final lift (mm)	Hold span (mm)	Arm RMSE	Hold pass
MuJoCo	65.08	5.165	1.128	No
Bullet	67.05	3.075	1.047	Yes
Gazebo	65.96	4.008	1.172	Yes
Drake	68.56	3.271	1.111	Yes
Isaac Sim	70.28	4.267	1.136	Yes
Newton	59.31	8.196	2.662	No

The successful Drake grasp provides a complete event chain: the dense native record first reports object contact at 5.3375 s, lift crosses 40 mm at 10.4 s, and terminal lift reaches 68.56 mm with 3.271 mm final span. Its 4k counterpart exceeds the final-hold stability threshold. The dense contact timestamp and the 15 Hz display have their respective sampling resolutions. The 20k arm-tracking RMS is 1.111 mrad. Across the six grasp cases, five pipelines have RMS 1.047–1.172 mrad, while the registered Newton pipeline has 2.662 mrad. This separates measured arm-following behavior from object-holding behavior: similarly small arm errors coexist with different object stability margins. The observed association localizes the comparison to sustained execution and contact; assigning its magnitude to one solver parameter additionally requires a matched-action parameter control.

Placement: location, opening, and withdrawal have different roles. All six terminal object planar errors are below 7.16 mm, within the shared 25 mm target tolerance. MuJoCo, Newton, and Isaac Sim satisfy the complete placement criterion. The Gazebo and Bullet cases have stable object placement and minimum tail-window gripper openings of 69.26 and 64.19 mm, respectively; their minimum tool–object distances are 39.52 and 39.62 mm against the 60 mm withdrawal requirement. Drake has a 10.49 mm final object span, a 24.95 mm minimum opening against the 55 mm release threshold, and a 35.50 mm tool–object distance. Its terminal planar error is only 3.93 mm. These records identify release/withdrawal and holding as distinct explanations for the task labels even when

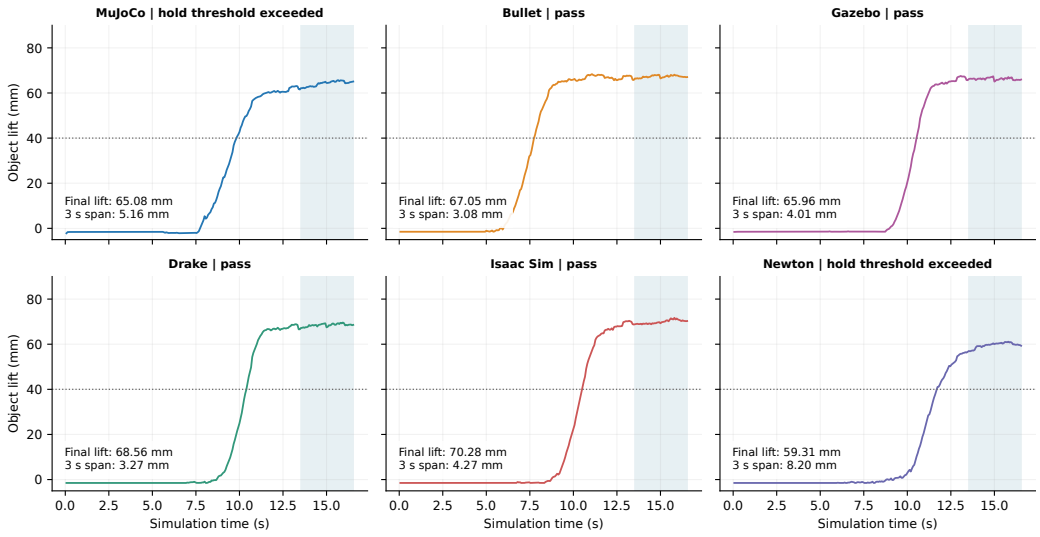

Figure S19: Native object-height traces during the six 20k grasping rollouts. Curves show direct 15 Hz subsampling of the recorded state at action times, without smoothing or time warping. The dashed level is the 40 mm lift target; shaded intervals mark the final 3 s. Stability values printed in each panel use the full archived evaluation window and all three position coordinates. The plots expose stage timing and lift evolution; the stability test additionally captures lateral motion.

the object reaches its target region. Distances are derived from measured joints through common-URDF forward kinematics and the native object center; the final window is evaluated jointly rather than at its endpoint alone.

Runtime and the scope of platform comparisons. After the first inference request, per-episode median model-service latency spans approximately 68.35–74.83 ms across these 18 rollouts. The simulation pauses during model inference, and eight actions cover about 0.533 s of simulation time. These timings describe the registered inference service and invocation context; physics throughput, render cost, and real-time closed-loop delay have separate measurement meanings. Within this dataset, the main platform-conditioned differences concern the learned commands and the stage-wise native response. Training wall time and model latency retain their hardware/load context.

F.4 GPT as assistance within the task-execution pipeline

GPT assistance is evaluated as a means of helping the trained policy complete its task. It receives the two images and task history, returns stage-conditioned text and an observation/continuation decision, and leaves numerical joint targets to the VLA. For the available matched aluminum/209001 cases, initial dual-camera images and scenario configurations agree. These cases use a separate scene/evaluation cohort from the aluminum/990001 training-budget comparison, with their own recorded model and execution identities.

The MuJoCo grasp pair finishes at 0.74 mm lift with pure VLA and 66.64 mm with GPT assistance; the final 3 s object span changes from 56.04 to 9.36 mm. This is an improvement in lifting and final holding behavior, with the full 5 mm stability threshold retained. The Bullet grasp pair finishes at 64.61/59.22 mm lift and 11.58/24.90 mm span for pure VLA/GPT assistance. A MuJoCo pushing pair changes from 0.91 to 28.73 mm axial error and reaches 78.34 mm vertical lift in the assisted rollout, showing a different action mode. The raw arm-target differences have RMS 0.0711 rad for MuJoCo grasping and 0.0854 rad for its pushing pair. GPT feedback is thus traceable through the generated actions and object response, with case-dependent completion effects. Initial-input matching controls the starting observation; policy sampling and subsequent closed-loop observations remain part of the paired execution. These examples characterize assistance behavior at their actual sample size.

The available stage-analysis snapshot contains 99 assisted rollouts and 18 pure-VLA scene-baseline rollouts, with unequal platform/task coverage. Its complete trajectories support selected process comparisons; platform-wide aggregate assistance estimates require matched denominators. The principal six-platform learning comparison remains the complete 18-case training-budget matrix above. A separate GPT direct-Cartesian-control run records successful lift and controlled release with eight model calls and 66.4 mm peak lift. Its action-generation role and endpoint criterion define a separate architecture-specific demonstration.

F.5 Cross-source diagnostics and evidence synthesis

The independently registered 72-episode process matrix crosses four source policies, six execution backends, and three tasks at a fixed seed. It uses the earlier 3,000-update protocol and a 32-dimensional representation with its registered action mapping. Command/nominal-path RMS is 96.30–129.17 mm and actual/command RMS 1.77–4.28 mm. It records zero approach events and zero contact in the 60 episodes with adequate active robot–object contact coverage. This cohort identifies an approach-stage separation. An independent fixed-checkpoint MuJoCo representation control, with matched initial conditions and random noise over three tasks and three seeds, reaches sampled contact in 8/9 episodes and full completion in 1/9 versus 0/9; the Wilson interval for 1/9 is [0.020, 0.435].

The separate 48-episode same-weight VLA/GPT cohort uses six platforms, four materials, two modes, and one scenario seed. Its actions run at 30 Hz with 240 Hz records and 960 Hz integration. All 48 tasks end before successful object contact. The 120 audited GPT responses contain 119 continuation decisions and one reobservation decision; the latter is an explicit model action. These approach-limited cohorts and the successful trained-policy cohort define different observed operating regimes and retain separate denominators.

Cross-platform finding and connection to the physical benchmark. Under the shared training and execution specifications, all six trained pipelines acquire successful pushing behavior, all six 20k grasp cases reach the lift target, and their final holding and release/withdrawal behavior differentiates task completion. Increasing the registered training budget coincides with improved validation prediction and additional successful grasp/place cases. Process evidence then distinguishes target generation, command realization, sustained contact, and release. This extends the physical benchmark's stage-conditioned finding into policy execution: similar target attainment can coexist with different holding stability or task-stage completion. Friction/support mechanisms are relevant during grasp holding, while separation and settling are relevant during placement. The local physical controls in the preceding appendices support those mechanism classes; the policy records supply the actual stages, commands, and responses on which their task relevance depends. The demonstrated value is a common, traceable basis for locating platform-conditioned learning and execution differences and prioritizing task-critical calibration. Real-task transfer remains a distinct endpoint requiring paired physical robot executions.

G Mechanical basis and native formulations

G.1 Force, friction, impact, and mode competition

The task design follows elementary equilibrium and motion relations. On an incline, $N = mg\cos\theta$ and tangential gravity is $mg\sin\theta$; Coulomb onset gives $\tan\theta_c = \mu_s$ when the assumed mode and quasi-static conditions hold. On a level plane, $P_s = \mu_s mg$ and the overturning balance $P_t h_p = mgb/2$ determine sliding/tipping competition. Force direction changes normal load and available friction; support width and application height change the moment balance. Rolling uses translation and rotation jointly, $v = r\omega$ for pure rolling, with moment of inertia describing filling and shape effects.

For a one-dimensional isolated impact, $m_1 v_1^- + m_2 v_2^- = m_1 v_1^+ + m_2 v_2^+$ and $e = -(v_2^+ - v_1^+)/(v_2^- - v_1^-)$ define momentum and restitution references. Sustained-contact, constrained, and multiple-impact cases retain the corresponding extra impulses and event windows. A complete body-force interval obeys the momentum residual in the main text. Gravitational potential plus kinetic energy

supports passivity and dissipation checks, with signed work from prescribed inputs retained where present.

G.2 Native contact and constraint representations

Generalized dynamics is $M\dot{v} + c = \tau + J_c^\top f$. MuJoCo's regularized constraint formulation can be expressed as a force optimization over its admissible set,

$$f = \arg\min_f \left[\tfrac{1}{2} f^\top (A + R) f + f^\top (a_u - a_r)\right], \qquad A = J_c M^{-1} J_c^\top,$$

with native regularization, reference acceleration, friction representation, and stopping rules (MuJoCo Developers, n.d.). Impulse methods update velocity through $v^+ = v^- + M^{-1} J_c^\top \lambda$, while platform-specific restitution, friction, and joint coupling determine the solved impulse. The used Bullet and PhysX paths retain their declared iteration and contact semantics (Coumans & Bai, n.d.; NVIDIA, n.d.).

An XPBD position-constraint update has the characteristic compliance term

$$\Delta\lambda = \frac{-C(q) - \widetilde{\alpha}\lambda}{\nabla C^\top M^{-1} \nabla C + \widetilde{\alpha}}, \qquad \widetilde{\alpha} = \alpha/(\Delta t)^2.$$

Native relaxation, body-pose reconstruction, restitution, and actual substep determine the complete algorithm (Macklin et al., 2016; Newton Developers, n.d.b;n). Drake SAP solves a discrete compliant-contact problem; hydroelastic contact further depends on geometry and pressure representation (Castro et al., 2023; Drake Developers, n.d.). Gazebo integrates the chosen backend's constraint assembly with its component execution and sampling phases (Gazebo Developers, n.d.; DART Developers, n.d.). The registered implementations and versioned source references are used to interpret the controlled evidence.

G.3 Deformable material and geometric quantities

Young's modulus E and Poisson ratio ν define Lamé parameters $\mu_E = E/[2(1+\nu)]$ and $\lambda_E = E\nu/[(1+\nu)(1-2\nu)]$ for the corresponding elastic parameterization. The deformation gradient $F = D_m D_r^{-1}$ connects current and reference tetrahedral edge matrices. Native models choose their stated constitutive energies and integration/contact treatments. Node displacement, signed element volume, and penetration are retained separately from derived material stress. This separation lets the common mesh/material-prior experiment describe numerical response and geometric support while preserving each model's physical assumptions.

H Asset fidelity, functional validity, and real-operation evidence

H.1 Evidence layers and the meaning of asset correctness

Asset validation connects a physical object to a specified computational representation. Four complementary layers are retained: visible-surface agreement, articulated assembly, effective collision/control behavior, and calibrated physical response. Each layer has its own observations and denominator. Geometric agreement constrains contact location and lever arms; articulation checks constrain the mapping from joint state to pose; functional tests establish that the declared geometry participates in the tested interaction; real observations characterize the corresponding operation and sensor response.

Table S11: Asset evidence and the conclusion supported by each measurement.

Layer	Observed evidence	Supported interpretation
Surface geometry	Registered surface distance, coverage, projected outlines	Agreement on the observed surface and residual location
Assembly/kinematics	Joint anchors, axes, limits, link transforms, URDF forward kinematics	Consistency of the articulated engineering representation
Physical function	Effective collision shapes, contact events, motion and tracking	Usability under the tested native configuration and trajectory
Real operation	Independent video, opening/closing stages, raw tactile channels	Observed timing and stage-specific hand response
Dynamic validity	Shared inertia/input, independent mechanical reference, native readback	Response of the stated physical model under a specified algorithm

These layers preserve the benchmark's causal order: asset and effective-input checks qualify the comparison; native contact/constraint and state records locate response differences; independent mechanics and controlled parameter changes support attribution. A rendered object, a geometrically matched object, and a dynamically validated model express different evidence claims. The Piper scan-calibration asset, the prepared room scene, the quantitative appliance mechanisms, and the archived policy assets retain separate identities.

H.2 Rigid registration, distance, and coverage

The appliance comparison sheets use rigid ICP with fixed scale and a 20 mm effective-overlap threshold. For a scan point set P and model surface $\mathcal{S}$, the aligned point is $T(p) = Rp + t$, with $R^T R = I$ and $\det R = 1$. A directional discrepancy is

$$d_{P\to\mathcal{S}}(p) = \min_{x\in\mathcal{S}} \|Rp + t - x\|_2.$$

The reverse direction evaluates sampled model points against the observed surface. For a declared retained-distance set D_δ, RMS and a tolerance fraction are

$$\mathrm{RMS}(D_\delta) = \sqrt{|D_\delta|^{-1} \sum_{d\in D_\delta} d^2}, \qquad p_\epsilon = |D_\delta|^{-1} \sum_{d\in D_\delta} \mathbf{1}(d \le \epsilon).$$

Directional coverage separately records the proportion of the source support that participates in the overlap. Sampling density, surface-area weighting, visibility masks, and trimming determine the precise denominator. The supplied sheets report bidirectional summary statistics; their retained summaries are reproduced here, with their original coverage labels. The sheets supply aggregate evidence, while point-level resampling and independent recalculation require the associated registration records. Fixed-scale alignment preserves size discrepancy while absorbing global translation and rotation. The 20 mm overlap threshold defines included correspondence support; the 5 mm fraction provides an additional accuracy description within that reported support.

The microwave sheet combines a physical photograph, three-dimensional discrepancy, and orthogonal projections. Its median of 2.52 mm and P95 of 10.18 mm describe a concentrated central distribution with a wider residual tail. The model-side 64.4% overlap and scan-side 100% overlap use different supports: all reported scan support can match a subset of a more complete model. The

Table S12: Reported appliance geometry evidence. All distances are millimeters. Coverage follows the directional labels of the supplied comparison sheets.

Quantity	Microwave	Breakfast appliance
Bidirectional median	2.52	5.05
Bidirectional RMS	4.75	8.59
P90 / P95	6.59 / 10.18	15.27 / 17.51
Reported fraction within 5 mm	84.1%	49.7%
Model-side overlap / local coverage	64.4%	18.0%
Scan / RGB-D point overlap	100.0%	100.0%
Effective-overlap threshold	20	20

breakfast appliance uses the visible surface of one RGB-D frame. Its 18.0% model-side coverage confines the geometric conclusion to that visible region; the 5.05 mm median and 17.51 mm P95 identify larger local residuals under this observation. Occlusion, pose, registration, depth observation, and model shape contribute to this residual space. Their individual contributions require corresponding independent controls. Different coverage and sensing protocols give each asset its own scope of interpretation.

H.3 Piper: shape, installation, and articulation

Section A.6 gives the complete numerical Piper comparison. The source data comprise two enabled whole-arm poses, one disabled whole-arm pose, and a separate clamp. The three scan-derived crossbar widths are consistent near 165.4 mm; the older geometry has a 145.0 mm crossbar. The newer template is closer in the observed crossbar region. Figure S22 provides spatial corroboration of the width diagnostic. Orange scan points also contain neighboring components and attachments, which explains the value of reporting the broad-region RMS alongside the local median.

The two enabled poses support a local finger-origin value of 137.5 mm compared with the older 135.8 mm definition. Fixed-point comparisons reduce visible-core RMS from 0.632–0.649 mm to 0.251–0.332 mm. These values characterize geometry/installation agreement using the registered scan regions. Full-arm encoder calibration, flange datum, internal mass distribution, and task TCP have separate measurement meanings. This separation assigns the observed gripper discrepancy to the asset layer before platform-level contact or control differences are interpreted.

Native import and functional evidence. A separately identified prepared Piper room asset was audited in Isaac Sim 5.1 at 480 Hz. Its URDF and USD eight moving joints agree in axes, anchors, limits, and stated drive limits; 205 visual meshes and 14 collision-source meshes have matching vertices. Maximum mass, COM, and inertia-element differences are approximately 4.29×10^{-8} kg, 4.81×10^{-9} m, and 1.58×10^{-9} kg m^2. Across 100 recorded states, forward kinematics from actual simulated joint angles agrees with native link poses within 0.000382 mm and 0.0000453°. These quantities establish representation/import consistency at the recorded precision.

The corresponding 38 s functional run retains finite states at 480 Hz, ground/table support, and appliance motion. A separate 4.5 s fixed-object gripper test records contact at both fingers, peak force magnitudes about 10.73/11.51 N, and release of contact on opening. These magnitudes come from interval impulse divided by step duration. The completed run reports effective collision shapes, while a separate rendered short run records 24 collision-mesh creation failures with GPU allocation errors. Run-specific effective-shape checks therefore qualify functional evidence. The archived scene also uses nominal 0.5/0.4 static/dynamic friction for several materials. Material-pair accuracy belongs to the measured-pair protocol, with that nominal scene kept as a separate configuration.

Tracking and geometry have distinct diagnostics. In the functional run, the drawer reaches a 146.637 mm peak for a 120 mm target before settling near 120.002 mm; the transient overshoot is about 22.2%. Joint-state forward-kinematic agreement simultaneously remains at the stated sub-micrometer scale. This pairing locates different questions at import geometry and driven response, respectively. It supports the benchmark's use of actual input, native state, and event-specific evidence rather than a single asset-level pass label. The prepared-scene audit applies to its identified Isaac configuration; the separate six-platform mechanism results retain the asset and controller definitions in Section D.

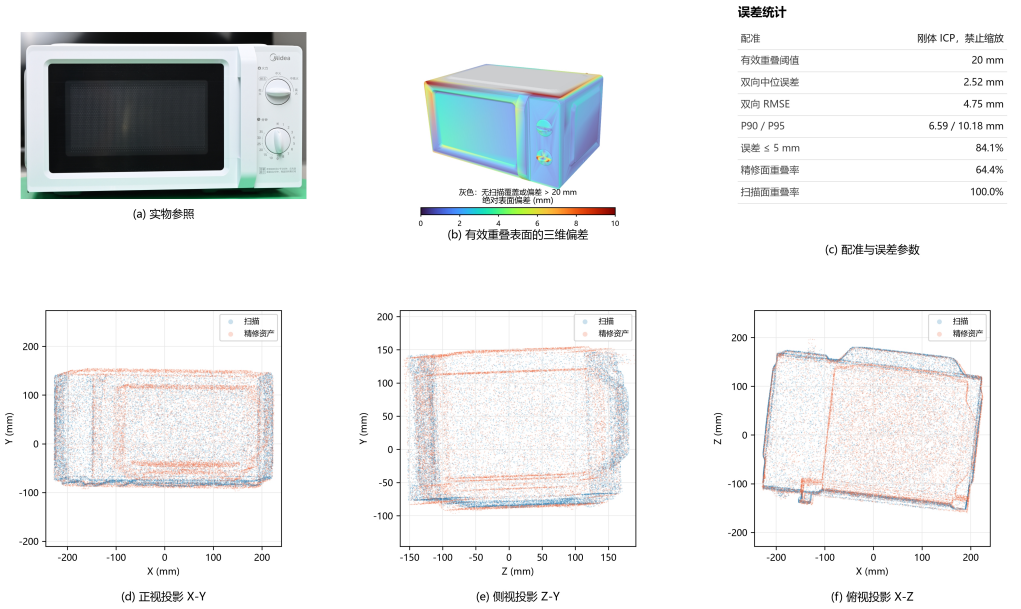

Figure S20: Microwave: physical reference (a), overlap-surface deviation (b), reported registration statistics (c), and X–Y, Z–Y, X–Z projections (d–f). Blue points denote the scan and orange points the refined asset. The color scale spans 0–10 mm; larger retained errors saturate that scale. Gray identifies uncovered surface or distance above 20 mm. The orthogonal views help localize silhouette and surface offsets after the same rigid alignment. Table S12 gives the numerical values in English.

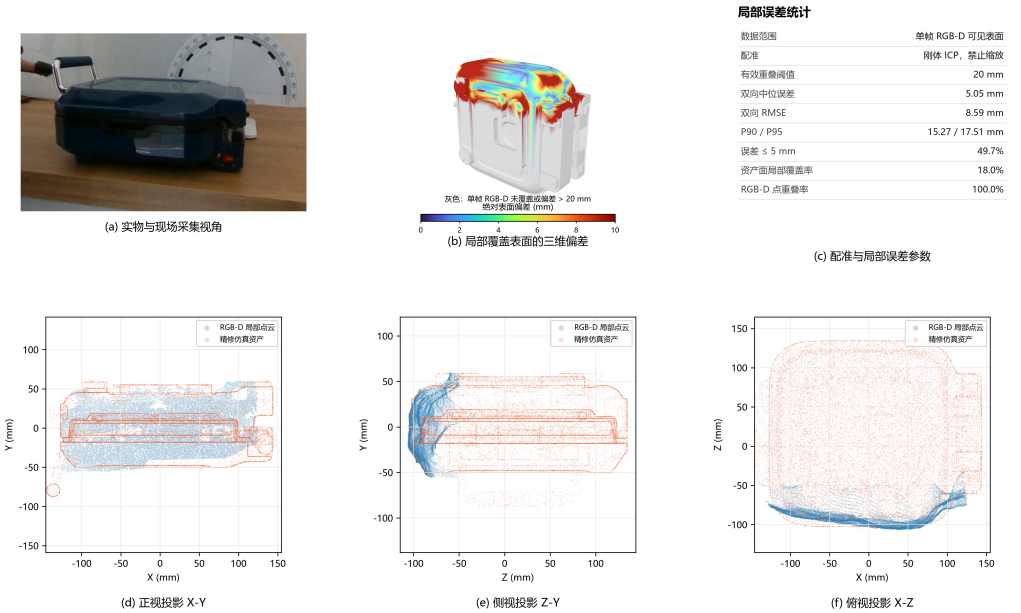

Figure S21: Breakfast appliance/electric griddle: physical capture view, local deviation, statistics, and three projections. Blue is the visible single-frame RGB-D support; orange is the refined asset. Gray indicates unobserved or above-threshold surface. The spatially restricted overlap, especially visible in the top projection, explains why the reported local error characterizes an observed part of the asset.

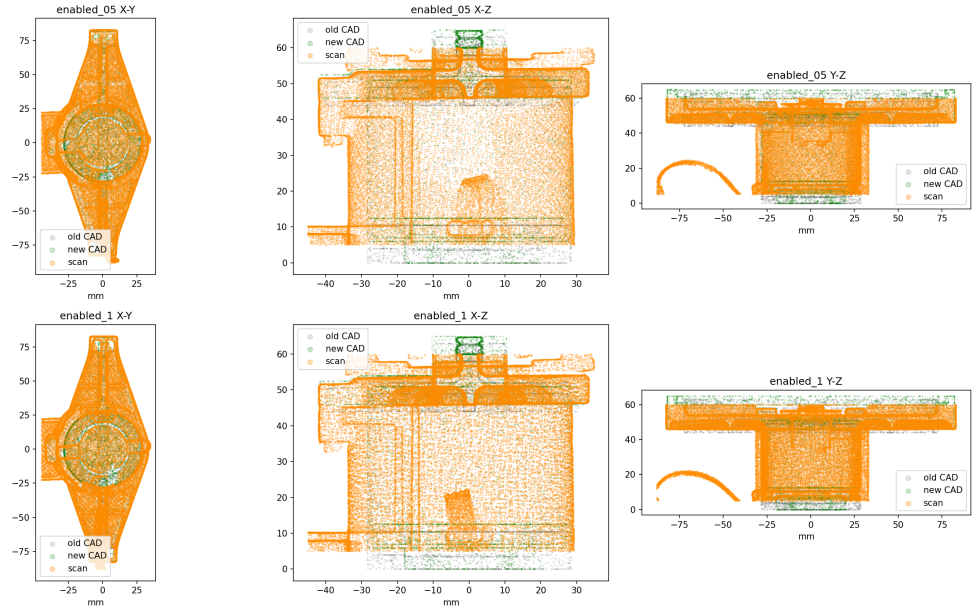

Figure S22: Piper gripper-base comparison in two enabled scan poses. Gray: older CAD; green: newer official template; orange: observed scan. Rows correspond to enabled poses 0.5 and 1; columns are X–Y, X–Z, and Y–Z projections in millimeters. The crossbar outline gives a repeated shape/version cue across poses. The projections retain attachments and broad-region residuals, complementing the quantitative local and wide-region results in Section A.6.

H.4 Real appliance trials: events, units, and visible variability

The two supplied real-operation panels each show five trials. Their row numbers are local figure identities. Each row combines a complete raw response, an opening peak image/time, an angle-observation frame, and a closing peak image/time. The images and glove stream use host-reception-time correspondence. Event-relative alignment compares local shape and phase variability, while within-row timestamps retain the original operation duration. U denotes the glove SDK's uncalibrated response unit; its meaning is distinct from a joint effort in N m and a calibrated contact force in N. A zero response at an angle-observation frame describes that displayed channel sample; the photograph independently identifies geometric phase.

Table S13: Displayed per-trial annotations. Each triple lists opening-peak time / angle-frame time / closing-peak time in seconds. Peak values retain unit U.

Asset	Trial	Time triple (s)	Opening peak (U)	Closing peak (U)
Microwave	1	5.27 / 9.31 / 16.15	29.62	22.45
	2	3.18 / 6.98 / 13.49	32.34	22.70
	3	2.66 / 6.26 / 12.43	17.58	20.61
	4	3.14 / 6.94 / 13.45	31.09	21.39
	5	3.57 / 7.24 / 13.52	24.96	20.64
Breakfast appliance	1	2.50 / 7.17 / 10.74	18.85	19.33
	2	3.81 / 7.28 / 10.52	18.84	17.82
	3	3.47 / 7.37 / 10.64	16.76	19.33
	4	3.33 / 7.53 / 11.60	14.39	17.10
	5	4.63 / 7.77 / 11.47	15.08	18.41

For the five displayed trial peaks, the microwave opening median/range is 29.62/[17.58,32.34] U and closing median/range is 21.39/[20.61,22.70] U. The corresponding breakfast-appliance values are 16.76/[14.39,18.85] and 18.41/[17.10,19.33] U. These are descriptive statistics of the annotations in Table S13. The microwave shows greater opening-peak spread than closing-peak spread in this

subset. The breakfast-appliance traces show several response lobes through the operation, and their local shape varies across trials. Hand contact, grip/support, and motion timing jointly enter the observed signal; separation of these contributions requires the matching channels.

The source panels also print summary annotations 27.64/21.82 U and 15.66/18.12 U. Those values differ from the medians of the displayed trial peaks. The present quantitative interpretation uses the explicitly defined five-trial statistics above. The source annotations remain visible for provenance and have an unspecified aggregation/window relative to this table. In general, peak of a pointwise median and median of trial peaks obey different operations:

$$\max_{\tau} \operatorname{median}_i r_i(\tau) \quad \text{and} \quad \operatorname{median}_i \max_{\tau} r_i(\tau).$$

Their difference illustrates the importance of naming the statistic; the exact computation behind the source summary labels requires its corresponding record. Five displayed trials describe the present figure subset; the separately registered ten-operation conditions in Section C keep their own identities and denominators.

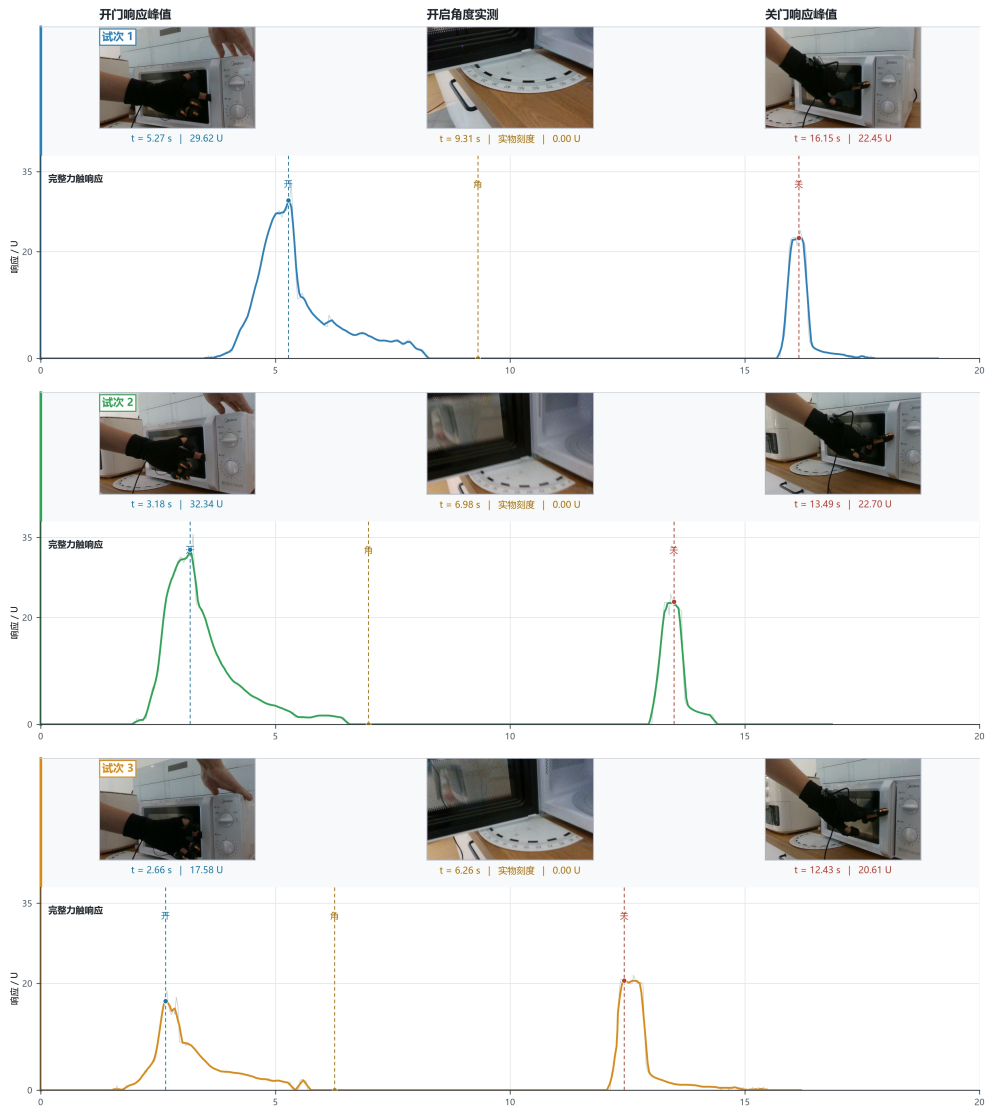

Figure S23: Microwave real operations, trials 1–3. Each row associates the opening response peak (left photo), angle-observation frame (middle), and closing peak (right) with the raw U-valued curve. Dashed lines mark the corresponding local times. The complete rows preserve operation duration and distinguish narrow closing pulses from the broader opening response. Original Chinese image annotations identify these three events and the trial number; numerical annotations are transcribed in Table S13.

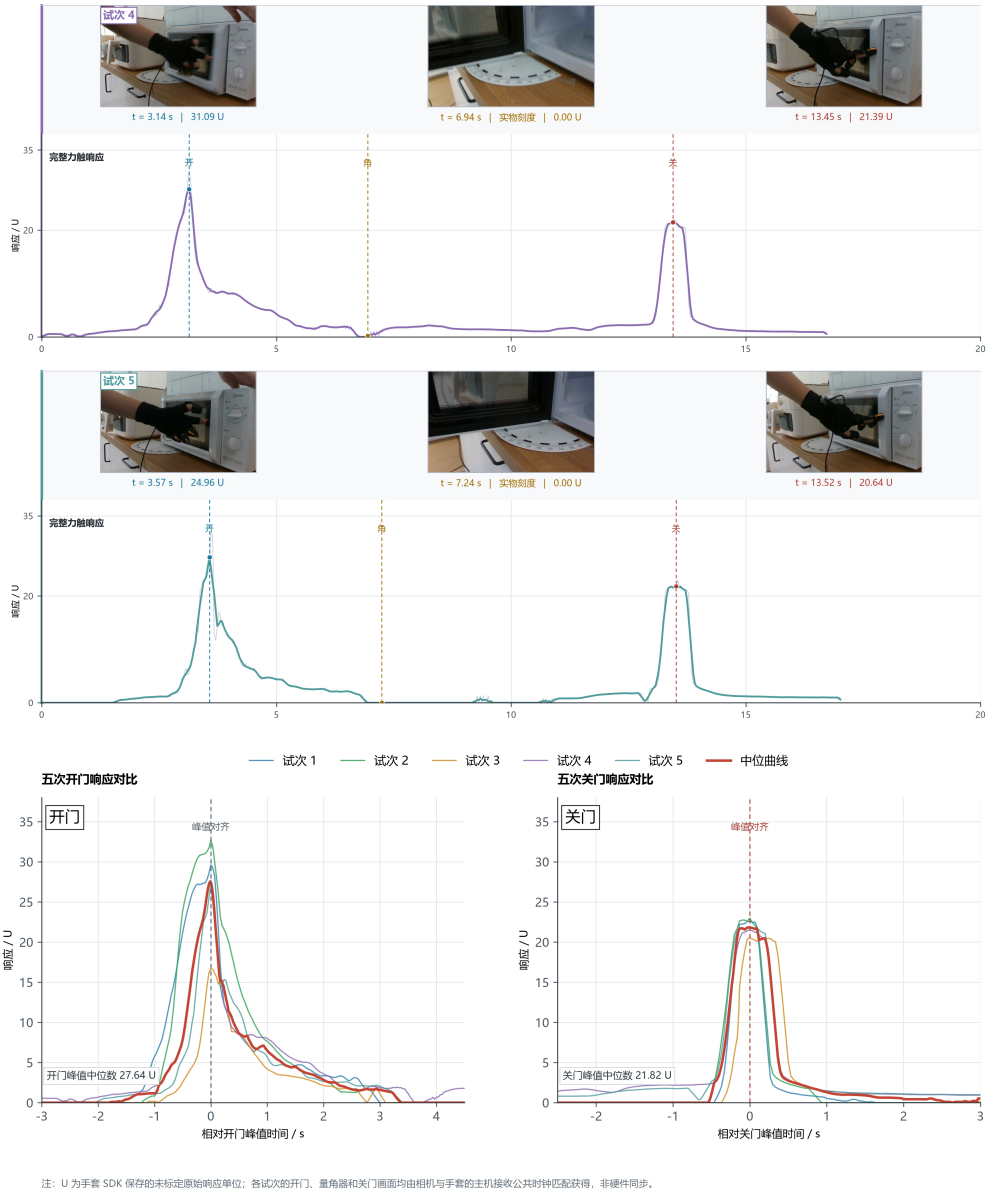

Figure S24: Microwave trials 4–5 and peak-aligned opening/closing comparisons. Colored traces are individual operations and the red curve is the displayed pointwise median trace. Alignment sets each respective peak time to zero and preserves amplitude in U. The source plot's summary labels retain their own aggregation status described in the text; the five-trial peak statistics use Table S13.



Figure S25: Breakfast-appliance real operations, trials 1–3. Opening, maximum-opening observation, and closure are paired with the continuous raw response. Several local lobes occur around lifting and closing, indicating that a single maximum summarizes only one aspect of the observed hand interaction. The figure preserves these lobes and the original phase timing.

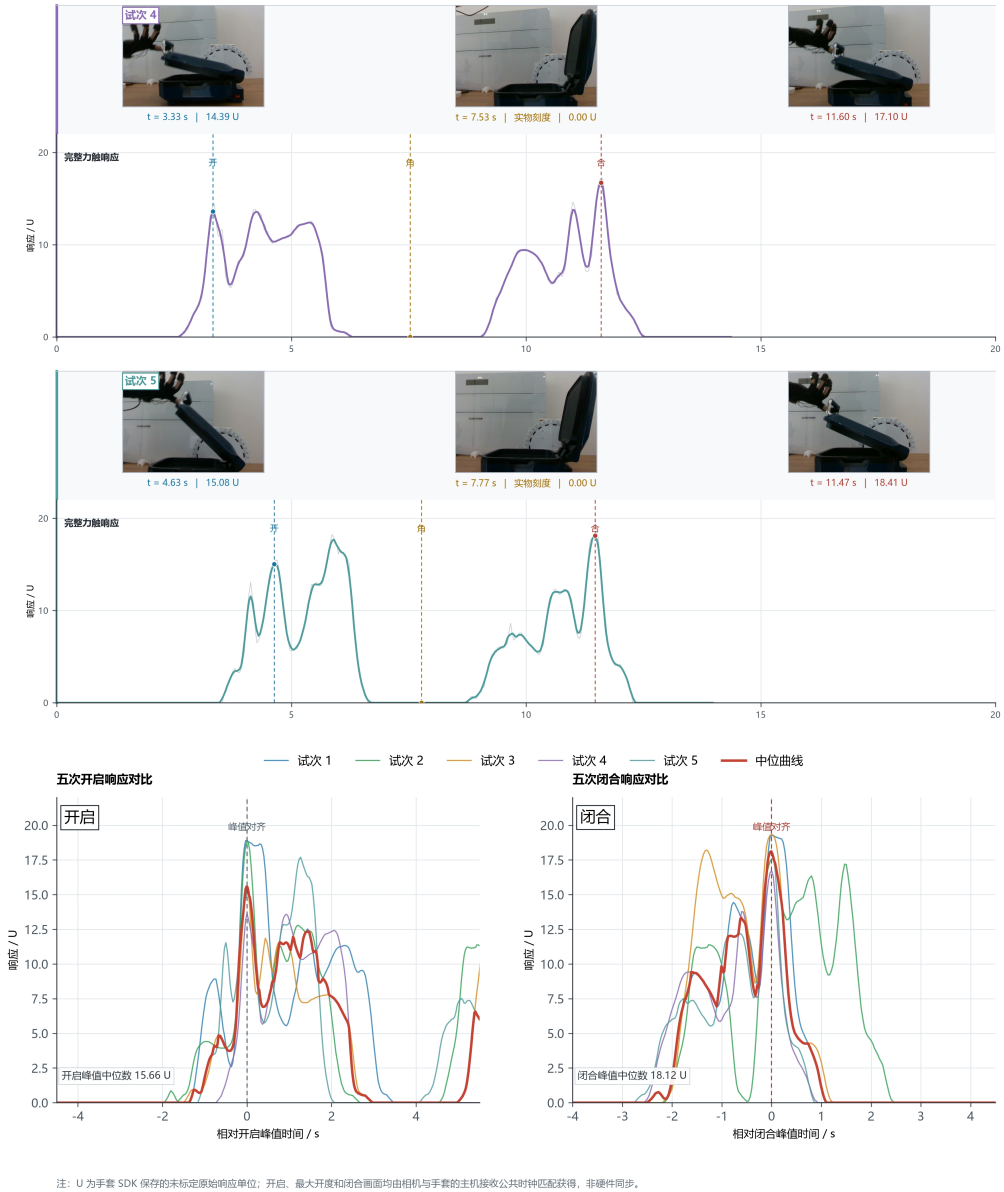

Figure S26: Breakfast-appliance trials 4–5 and aligned opening/closing response. Peak-relative alignment supports comparisons of local duration, lobe structure, and amplitude across five operations. Photos preserve the geometric context. The same raw-unit and source-summary distinctions used for the microwave apply here.

H.5 Independent validation and final evidence synthesis

The evidence separates construction inputs from evaluation outputs. Scan-based fitting and same-scan residuals establish internal geometric consistency, with cross-pose comparisons adding robustness to pose variation. Independent real photographs and operating sequences add observed object and task identity. Real hand data provide phase-conditioned observations. The independent eccentric-hinge recurrence and its controlled torque/inertia provide the external dynamic reference for the six-platform articulation comparison. A replayed angle is a prescribed boundary input; its tracking error evaluates the controlled response to that input. Physical torque accuracy uses a separately specified dynamic reference and calibrated measurement semantics.

Table S14: Final asset-to-response evidence chain and scope.

Object / test	Mutually supporting evidence	Conclusion scope
Piper scan asset	Repeated crossbar outline, fixed-region finger residuals, local installation fit	Version/assembly discrepancy localized at the observed gripper region
Prepared Piper scene	URDF/USD readback, runtime FK, effective-shape/contact and tracking records	Import consistency and trajectory-specific physical functionality
Microwave surface	Photograph, bidirectional distances, 64.4% model-side overlap, projections	Quantified agreement over the reported observed support
Breakfast surface	Single RGB-D view, 18.0% model coverage, local error distribution	Localized surface evidence with explicit visibility scope
Real opening/closing	Photo-linked timing, five-trial raw curves and peak distribution	Stage-specific hand response and observed trial variability
Six-platform hinge	Common inertia/torque, native readback, independent recurrence, relaxation controls	Configuration-specific dynamic mechanism described in Section D.3

Together, these comparisons give asset validity a testable, layered meaning. Geometric residuals are localized and coverage-qualified; articulated inputs are checked at the native interface; response discrepancies are assigned to the first separating physical or observation stage. The real operation panels corroborate phase identity and variability, while the independent mechanical controls support the numerical mechanism. This evidence structure links asset construction to the benchmark's physical conclusions without merging geometry, controller response, and hand-sensor intensity into one accuracy quantity.