

Supplementary Material for MOTIO

SpatialtemporalAI

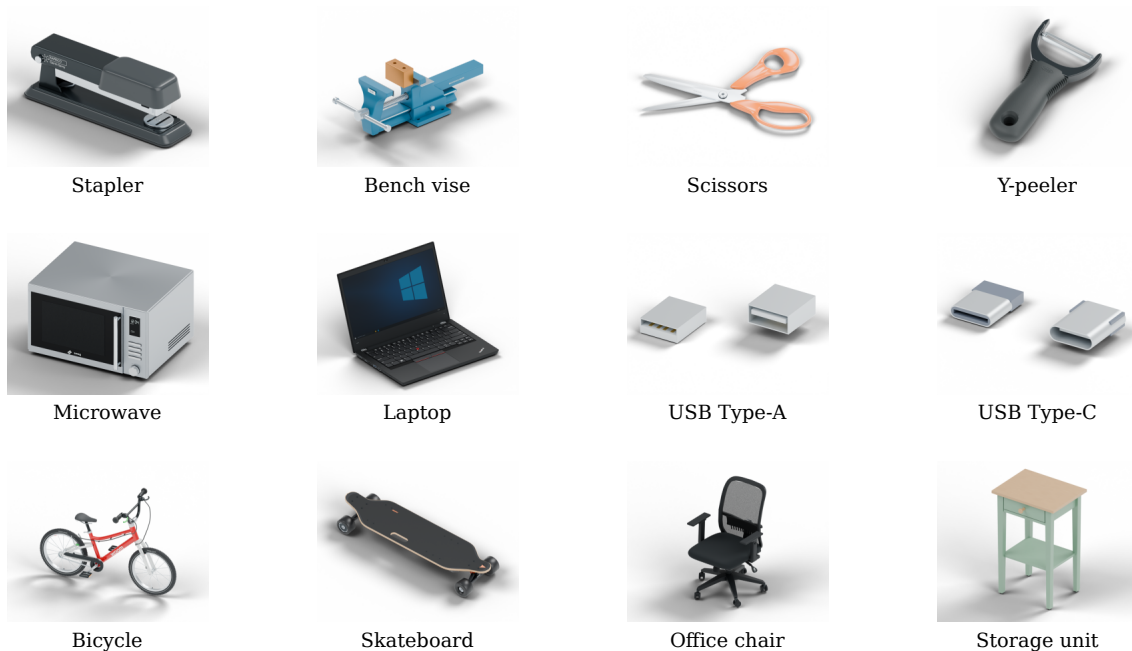
S1. Agentic 3D assets and task-executable physical representations

S1.1. Asset production and library units

Agentic 3D begins with a product or environment specification, searches available product reference images and manufacturer dimensions, and constructs an editable model with the intended scale and component structure. Most models are authored with Blender scripts and interactive operations; FreeCAD supports products whose geometry is driven by parametric constraints, drawings and engineering interface dimensions. Human review establishes the version that is passed on to subsequent task preparation. Source visual meshes, part transforms, component identifiers and metric frames remain accessible during simulation-model construction.

The asset census is defined at *family* level. The local msj-3d library contains **78 asset families** (73 product/paired kits, 3 scenes and 2 hardware assemblies); simulation task assets span **104 families**; **59 families** appear in both catalogues; their union contains **123 families**. Historical revisions, repeated instances and mesh components of one family are grouped rather than counted as independent products or tasks. The local library contains 71 Blender-only families, five FreeCAD-only and two with both representations. These counts describe the asset substrate, while task acceptance is registered separately at the executed-skill level.

Agentic 3D: reviewed assets for executable interactions



Blender + FreeCAD · scale, structure and articulation

Figure S1: Reviewed Agentic 3D families across tools, engineered interfaces, product mechanisms and office furnishings.

S1.2. From reviewed product to physical contact model

An accepted product provides visible geometry, dimensions and component hierarchy. The task-ready physical model additionally assigns collision proxies, joints, passive constraints, contact sites and event measurements.

Its robot-facing interface exposes named object frames and phase targets, which the coding agent uses to construct arm and hand references. Human asset acceptance evaluates dimensions, geometry and component relationships; the skill-acceptance event evaluates actual executed task outcome and motion quality.

Rapesco stapler. The accepted `Rapesco_Marlin_Stapler.blend` asset is approximately $159 \times 41 \times 53$ mm. Its visible base, lever and magazine meshes follow corresponding rigid bodies while separate collision proxies represent the press cap, driver, anvil and paper-support surface. Lever and magazine rotate about local y axes against task-model springs (lever range 0–0.125 rad, spring 18; magazine range 0–0.075 rad, spring 1.5). Driver travel is coupled to lever rotation through a one-way lost-travel mechanism. The task starts from three supported rigid A5 paper sheets. A gated staple event requires a driver/press/contact configuration sustained for **12 steps at 400 Hz (30 ms)**, together with right-palm loading and left-hand paper stabilization; the event activates a paper connection proxy. The program uses press-cap/driver/paper frames to generate the right-palm normal and left-hand support targets, followed by unloading and withdrawal.

HEUER vise. The HEUER 100 product references include a 100 mm jaw width, 125 mm maximum opening and 50 mm clamping depth. Its reviewed scene separates a fixed seat, sliding jaw, screw, translating handle and initially supported workpiece. The jaw slides along a local x axis while screw rotation is coupled to jaw displacement by a **4 mm/turn task-model lead**. The handle slides along its local y axis within ± 72 mm. The skill uses screw angle to generate the right-palm handle arc and regrasp targets and checks jaw opening, opposing clamp contact, workpiece motion and retention after hand and support withdrawal. These model interfaces demonstrate how Agentic 3D feeds robot-native object targets.

Table S1: Task-facing asset to physics interface examples.

Product asset	Physical model	Skill interface / event
Rapesco 159×41×53 mm	Base, lever, magazine, press cap, driver, anvil and contact proxies; articulated lever and magazine.	Right-palm press-cap target, left-hand paper stabilization, contact-gated paper connection and withdrawal.
HEUER vise	Fixed seat, translating jaw, screw, moving handle, supported workpiece; screw/jaw coupled at model lead 4 mm/turn.	Hand arc referenced to screw angle; opposing clamp, jaw clearance and retained workpiece measured at task completion.

S2. Human demonstration as staged contact context

S2.1. Recovered reference observations

The bench-vise development record contains an initial natural-language request to implement clamping by turning a vise handle while searching for a human operation video. The coding agent inspected two actual reference videos and two corresponding contact-sheet views during the original session (recorded image views at **2026-10-13 04:51:21 UTC** and **04:52:14 UTC**). The steel vise clip illustrates workpiece placement, handle turning and release; the supplementary Ono Kosuki/Pexels wooden-wheel vise clip supports the stabilize–turn–regrasp–release interpretation. The latter includes a supported workpiece at ~ 0.5 s and changes of hand position around the wheel at ~ 2.5 s and ~ 4.8 s. These are observations used as *task context*; robot-specific geometric and control targets come from the accepted asset and executable program. Main Figure 4 shows an illustrative Virtual Guru/YouTube third-person reference; the original agent context contains the steel-vise and Ono Kosuki/Pexels observations described above.

S2.2. Accepted Case 43 staged program

For v0.43.0/adaptive_full101, the final 48-second program contains right-hand reach (0–6 s) and close (6–8 s), then four seven-second half-turn cycles from 8–36 s. Within each cycle the phases are turn ($u=0$ –3.5 s), unflex (3.5–4.4 s), retreat (4.4–5.3 s), reposition (5.3–6.3 s), and regrasp (6.3–7.0 s), yielding three

intermediate regrips. `goal()` constructs a wrist/palm arc around the handle and a grip/retreat clearance target while the screw angle updates the sliding-jaw location. Final tighten occurs at 36–37 s, right-hand release at 37–39 s, workpiece-support withdrawal at 39–41 s, and an object-retention check at 41–48 s. The accepted rollout completes **2.052235 handle turns** and narrows the jaw opening from **38.000 to 29.987218 mm**, retaining the workpiece after withdrawal.

Table S2 summarizes the accepted right-hand stages and physical checks.

S2.3. Additional video-guided tool example

A separate screwdriver history includes browser playback of Helpful DIY (ratcheting screwdriver at approximately 1:20) and Sikana (basic screwdriver operation around 1:55–2:05). Observed intent covers a supported axis, wrap grasp, back-and-forth ratchet motion and axial pressure. The agent’s robot-native task uses a supported horizontal tool axis and an $M5 \times 0.8$ screw mechanism. This example shows how the agent uses video-derived grasp and motion intent to define robot-native tool interactions.

Table S2: Case 43 right-hand task stages and physical checks.

Window	Program phase	Observation
0–8 s	Reach and grasp handle bar; workpiece initially supported.	Initial reach/contact.
8–36 s	Four 7-s half-turn cycles with unflex, retreat, reposition, regrasp.	Rotation, contact and jaw travel.
36–41 s	Final tighten, release hand and withdraw support.	Clamp loading and release.
41–48 s	Hold test after both hand and support withdraw.	Workpiece retention and gap (38 to 29.987 mm).

S3. Execution-grounded skill revision and control implementation

S3.1. Stapler program changes following human feedback

The full107 stapler candidate was reviewed with explicit feedback that the press should be driven by the palm rather than fingertips and that the withdrawal should look more natural. The accepted `palm_final10` version includes `palm_press` (13–16 s), `hold_stroke` (16–17 s), `unload_palm` (17–19 s), `release_papers` (19–21 s), `elbows_retract` (21–24 s), `lower_arms` (24–29 s) and final completion (29–31 s). The right palm normal aligns with the press-cap contact surface, while the left hand maintains paper stabilization. Reference corrections use lever-angle error around an approximately 0.108-rad target to adjust palm height; after the staple event the program briefly maintains the palm position before unloading. Retraction includes clearance to the tabletop and return toward the terminal arm posture.

The next 31-second physics rollout records a `clinch_staple_proxy` event at **14.8125 s**, with a right-palm pad/press-cap contact force around **14.8893 N** at the gate and the supporting left hand in contact. The trigger requires **30 ms** of sustained event satisfaction. After the event, the mechanism returns passively, the paper remains connected and supported by the table, and both hands withdraw. The human reviewer accepts this version as `v0.47.0/palm_final10`.

Table S3: Rapesco feedback-driven program change, Case 47.

Review input	Program revision	Physical result
Palm-driven press and natural hand withdrawal.	13–16 s palm press, 16–17 hold, 17–19 unload, 19–21 paper release, 21–24 elbow retract, 24–29 lower arms.	Staple connection event 14.8125 s; paper joined, hands withdrawn by 31 s; human approval.

S3.2. Physical execution and specialized material simulation

The common rigid-body and articulated-object backend is MuJoCo (Todorov et al., 2012). Whole-body standing and locomotion references use frozen ScaleBFM, with task-specific inverse kinematics, Revo2 hand joint targets and bounded actual-state contact feedback. For a supported or seated precision task, direct gravity-compensated joint control and local IK can be used. Mango-peeling Case 40 couples an accepted MuJoCo robot trajectory to an Isaac Sim 5.1 PhysX deformable fruit-peel representation. Additional LeHome-based tools supply project-specific deformable simulation capabilities. The accepted G1-OPT mango-peeling execution supplies the material-interaction example.

S4. Office reconstruction, layout variation and Real2Sim2Real

S4.1. Metric scene and localization

The appearance model uses 3D Gaussian splatting (Kerbl et al., 2023). The reconstructed office spans three connected rooms; Room 2 is approximately **29.3 m²**. The scene corpus includes an appearance reconstruction of roughly **1.86 million Gaussians**, a sparse colored point cloud of roughly **1.17 million points**, and independent collision, navigation and interaction representations. Task and scan coordinates are aligned to measured room geometry. A G1-mounted Livox MID-360 localizes against the mapped room; head D435i RGB-D and an external SAM3 service provide bottle/table observations, while the bin pose is taken from earlier real measurements. Thor plans body and Revo2 hand references for the measured layout and checks them in MuJoCo before loading references to the running ScaleBFM M Mode 7 controller with absolute-position feedback.

S4.2. Full-task layouts (31 runs)

Eleven named layouts consist of eight single-factor changes and rear/left/right table placements; twelve development-random and eight post-freeze random layouts complete the set of 31. The latter eight use seed 2026092301. Candidate positions are sampled within task-frame bounds: table $(x,y) \in [0.7,1.4] \times [-0.25,0.75]$ m, bottle relative to table $\in [-0.27,0.27] \times [-0.37,0.37]$ m, and bin $\in [-1.2,-0.25] \times [-1.3,1.0]$ m. Candidates with table/bin overlap, bottle outside tabletop support, or no feasible stance/path are rejected. Each retained layout triggers regeneration of stance, approach angle, avoidance route and phased body/hand targets. The photo-corrected simulation completes **31/31** full-task executions with continuous carrying grasp, release and return to stand, with durations **61.2–138.6 seconds**.

S4.3. Hand-fixture object properties and paired revision

The 44 distinct bottle property configurations vary diameter **58–68 mm**, height **220–235 mm**, and mass **20–600 g**, with mass and inertia updated with geometry. The constructed set combines 27 grid points and additional sampled/named settings, including twelve Latin-hypercube samples and five additional unique named settings after deduplication. The LHS generator uses $n=12$, seed 20260923, random offsets in twelve strata per axis and permutation before mapping to each physical range. The short hand fixture checks opposing thumb-plus-two-finger contact during **8 < t < 10 seconds**, then release. The accepted setting passes **44/44** short window tests.

The before/after contact refinement is paired on the **same twelve LHS settings**, included in the 44. Earlier configuration holds through the complete window in 11/12 settings (mean 95.83% window grasp retention); revised configuration holds in 12/12 (100%). Setting 01 rises from 50% to 100% grasp retention, while the other eleven paired settings remain at 100% in both versions. The main paper summarizes these paired outcomes in Table 1; Figure S2 presents the complete layout and object-parameter coverage.

S4.4. Three perception-conditioned whole-body sequences

The three chosen full-body conditions are **measured** (64×227 mm, 600 g), **narrow** (58×220 mm, 600 g) and **wide-empty** (68×235 mm, 20 g); their simulated durations are **108/112/108 seconds**. Each completes all

eight stages with vision-conditioned bottle estimation, carried-object grasp, release and return to stand.

S4.5. Physical deployment and source videos

One archived site-layout validation completes eight stages in a **124-second** simulated task, including **171 seconds** of planning-and-validation wall-clock time and bottle placement within **4.2 cm** of the bin center. The accepted physical real05 is a **101.7543-second** continuous grasp–carry–dispose execution without runtime retry. The corresponding accepted Case 9 vision_bottle_to_bin v0.9.0/full_v4 is a separately recorded **112-second** MuJoCo rollout. The real04 door-cart sequence provides another **98.41-second** long-horizon physical episode. All source durations and task-phase correspondences follow the actual archived media.

Simulation-side variation: layouts and object parameters

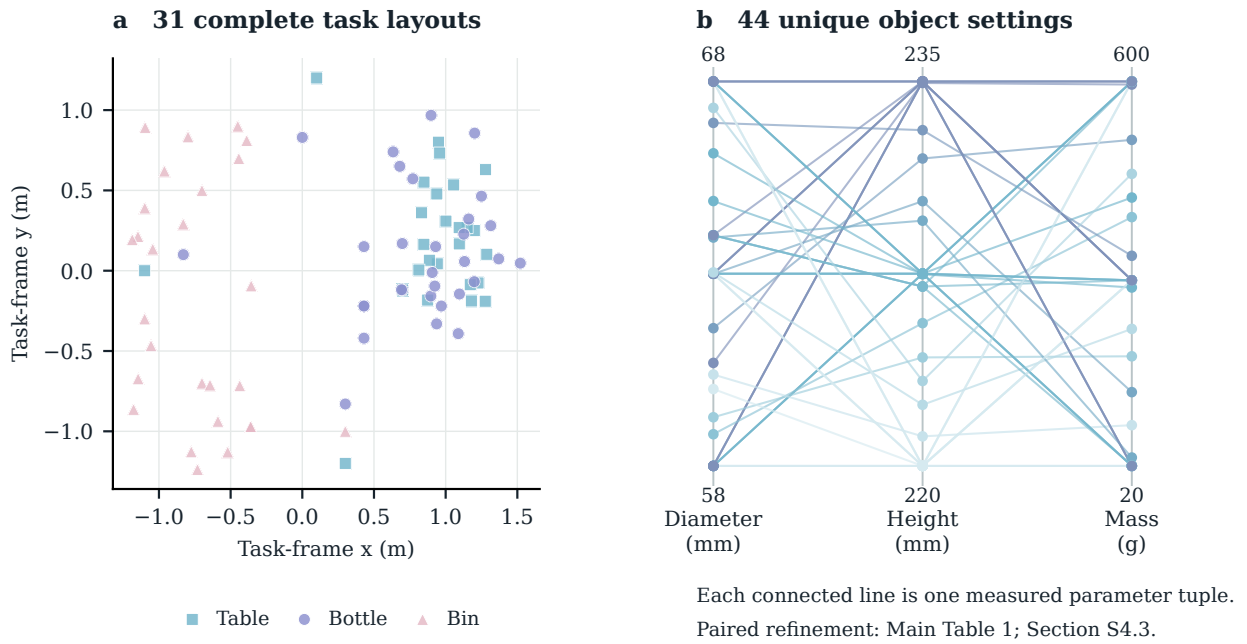


Figure S2: Detailed simulation-side variation. Left: positions of table, bottle and bin for 31 full-task layouts. Right: 44 unique simulated bottle-parameter settings across diameter, height and mass; each connected line represents one tested setting.

Table S4: Bottle-to-bin simulation variation units and completion checks.

Study	Variation and criterion	Observed result
Room layout	11 named, 12 development-random, 8 post-freeze; full grasp, carry, bin release, return.	31/31 full-task completions; 61.2–138.6 s.
Bottle properties	58–68 mm diameter, 220–235 mm height, 20–600 g; hand fixture at 8–10 s.	44/44 opposing-grasp windows.
Paired refinement	The same 12 LHS settings, earlier versus refined contact window.	11/12 to 12/12 full-window retention.
Visual whole-body	Measured, narrow-heavy and wide-light parameter conditions.	3/3 eight-stage completions.

S5. Quantitative skill-development histories

S5.1. Cohort and measurements

The consecutive registration set comprises Cases 40–73 (34 skills). Thirty-three have precise task-request and first human-approval timestamps. One skill is the unit of analysis; the agent-turn time is the union of open-run intervals including tool and resource waits; calendar time is task request to first approval including gaps between turns and human review. Retained dynamic versions are content-distinct full-body state trajectories; visual inspections are actual image/frame-review calls. The corresponding medians and IQRs for first approval are: **97.0 min** (42.5–170.3) agent-turn, **2.80 h** (1.33–9.80) calendar; **2 turns** (1–2); **18 versions** (7–31); **10 visual inspections** (7–18); **1 substantive human revision prompt** (0–1). The 32 later-presentation-aligned task histories have median accumulated agent-turn **123.3 min** and median elapsed calendar **3.89 h**.

The empirical cumulative distributions show **25/33** first approvals at or below 3 hours recorded agent-turn, **21/33** at or below 20 retained versions, and **17/33** at or below ten inspections. These are distinct thresholds for a common 33-skill first-approval cohort.

Table S5: First human-approval development records for 33 numbered skills. Agent time in minutes, calendar time in hours.

ID	Skill	Agent	Calendar	Versions	Visual
40	Mango peeling	80.6	1.67	9	17
41	Caliper measurement	204.2	18.52	7	18
43	Vise clamping	93.8	19.51	5	6
44	Drawer retrieval	170.3	19.53	8	13
45	Dowel insertion	79.3	19.48	3	9
46	Gluing paper	123.2	2.42	31	20
47	Stapling	164.1	2.89	18	8
48	Phone/WeChat	124.3	2.23	12	24
49	Printing/retrieval	576.1	9.80	58	16
50	Walking/sweeping	406.5	6.98	47	29
51	Water dispenser	139.1	2.79	18	17
52	HDMI insertion	126.4	2.58	19	15
53	Presentation remote	207.8	10.92	19	13
54	Stamping	46.1	1.33	13	7
55	RGB-D bottle pickup	22.6	0.40	6	3
56	Drinks delivery	143.2	2.54	46	19
57	Wire cutting	31.4	2.15	8	8
58	Airbrush painting	25.0	0.44	4	7
59	Microwave loading	221.0	4.13	34	18
60	Refrigerator storage	234.6	4.12	64	10
61	Microwave cycle	231.3	7.18	32	22
62	Hairdryer	31.6	0.55	7	10
63	Standing fan	18.1	0.32	5	5
64	Bicycle pump	45.9	0.78	19	5
65	Motorbike riding	25.6	0.57	16	8
66	Two-G1 sofa carry	57.3	0.99	22	8
67	Escalator	89.7	1.54	21	13
68	Sliding window	97.0	3.66	24	9
69	Backpack zipper	42.5	8.43	17	7
70	Fishing	228.8	13.16	48	26
71	Electric skateboard	11.5	0.20	6	6
72	Flexible takeaway bag	115.7	12.18	34	19
73	Screw tightening	29.6	172.42	5	7

Time to the first human-approved skill

33 timestamped skills from Cases 40–73



One dot per skill. Agent-turn time includes tool execution and resource waits.

Figure S3: Time to first human approval for 33 matched skills. Left: recorded agent-turn minutes; right: request-to-approval calendar hours on separate logarithmic axes.

Human-guided refinement expands accepted skills

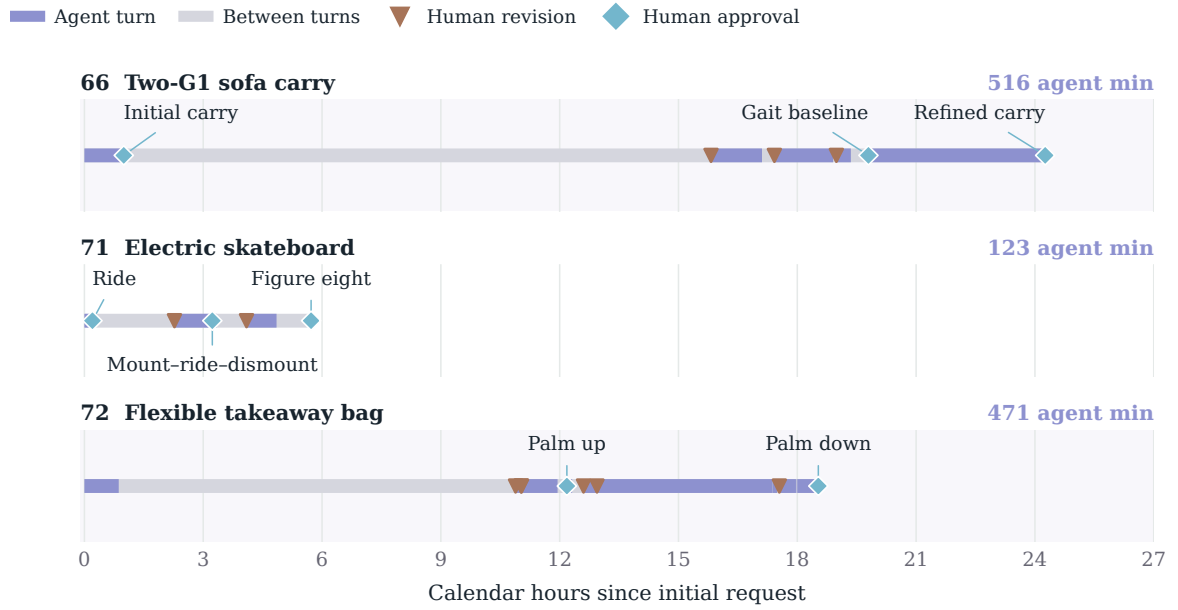


Figure S4: Recorded human-guided refinement and approval milestones for Cases 66, 71 and 72: two-humanoid sofa transport, electric skateboard and flexible-bag manipulation.

S5.2. Screw-tightening candidate outcomes

The first candidate has maximum thread residual **269.44 μm** and final screw advance **0.7697 mm**. The accepted candidate reaches **64.58 μm** (76.0% reduction) and **0.8019 mm**, against a target of **0.8 mm**, with peak unintended contact decreasing from **23.95 N to 0 N**. Full candidate states full01 through full04, including any early-ended trace, are retained in the supporting measured outcomes series.

Screw-tightening development across retained candidates

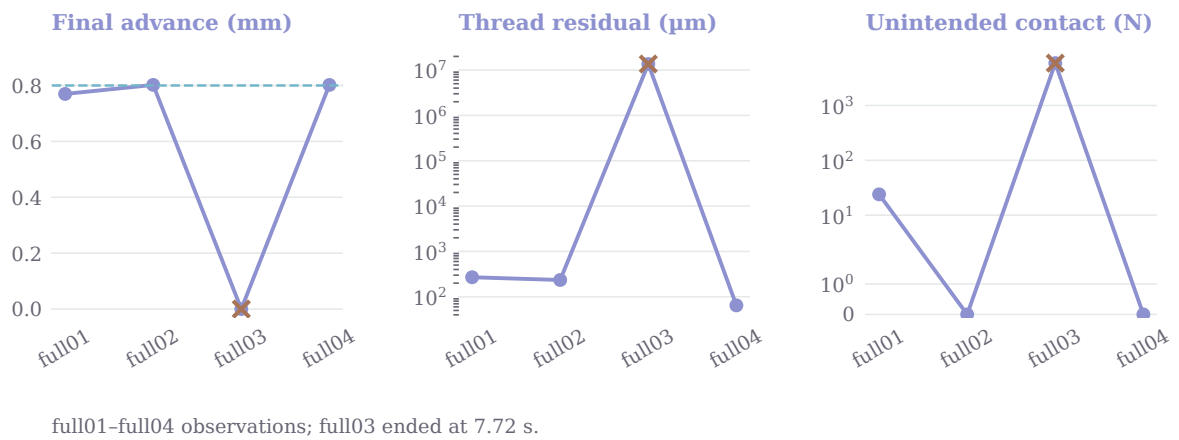


Figure S5: Complete measured screw-tightening candidate sequence full01–full04, including the early-ended full03 trace.

S6. BC/Dagger training and execution interface

S6.1. Input and reference target

The accepted checkpoint contains **3,127,287 parameters** and predicts a horizon of six future references sampled every **20 ms**, from t to $t+0.10$ s. Each 59-dimensional prediction combines **29 body-joint targets**, **12 hand-joint targets**, **six palm position components** for two palms, and **12 palm 6D rotation components**. The model consumes a **176-slot task/robot observation** containing skill and role indicators (20 and five categories), phase progress and public clock, root and palm states and explicit geometric memory, together with **32 mechanism values and their 32-bit validity mask** and task/phase encodings. The final configuration has `reference_history=False`: reference generation uses current publicly tracked phase and encoded geometric memory, while the low-level frozen tracker continues its own native proprioceptive history handling.

A 50 Hz task loop decodes the six-frame target block into robot body, hand and palm references and interpolates between samples. Frozen ScaleBFM tracks the body; a bounded actual-state arm IK and hand/contact-feedback executor applies stage-specific corrections. Object-frame codecs and public event gates are shared with the task program.

Table S6: Final BC/Dagger student observation and reference interface.

Component	Frozen accepted configuration
Observation	176 task/robot state slots, 32 mechanism values with 32 validity-mask bits, skill and phase identifiers, public clock and explicit geometry memory.
History	<code>reference_history=False</code> ; current public phase and geometry memory condition reference prediction.
Output	Six references $t, t + 0.02, \dots, t + 0.10$ s, each 59 channels: 29 body, 12 hand, 6 palm positions and 12 palm 6D rotations.
Executor	50 Hz reference decoder, frozen ScaleBFM tracker, bounded arm IK and task-event/contact feedback.
Training	BC initialization, student-state Dagger expert labeling and event-gated recovery corrections.

S6.2. Expert labeling, recovery and observed training stages

BC initializes the student using references produced by agent-authored expert programs and their simulated executions. Dagger applies the expert to student-visited states to create fresh labels; event gates may hold and requery a reference when an expected physical state transition has not occurred. Example gates include waiting for a staple event, caliper jaw contact, and scissors seam completion. Recovery labels reflect actual observed task state and task-program correction criteria.

The ten-task extension aggregation set contains **63 episodes and 92,130 frames**, of which **77,130** have `expert_executed=False` and **142** within the total carry a recovery flag; Student-execution and recovery-flag counts index the same collection. Across successive *joint system-development checkpoints*, the ten-task completion trace is BC 2/10, D1 5/10, D2 7/10, D3 8/10, D4 8/10, D6 9/10 and D7 10/10, with representation, reference-prediction and execution-interface improvements alongside data aggregation.

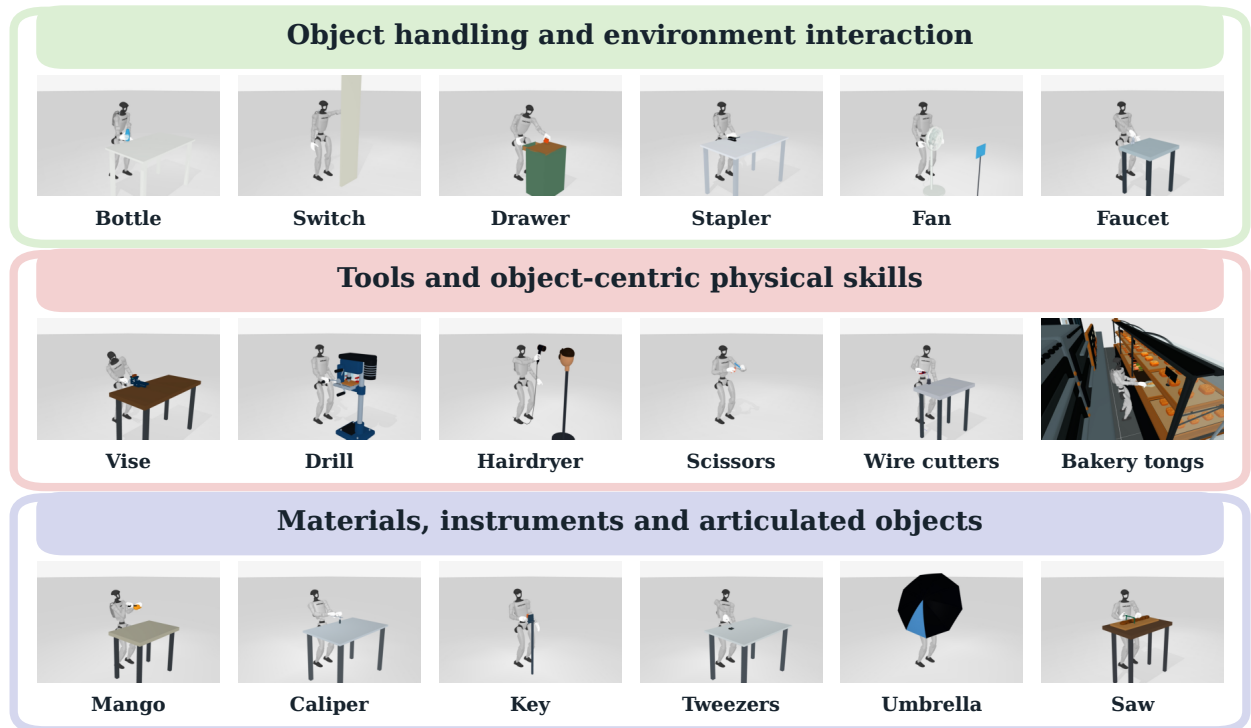
S6.3. Fixed suite and task galleries

One accepted checkpoint executes a suite of 18 fixed simulated tasks with **18/18** completion on recorded development runs and **18/18** on same-condition replay; inference involves **zero expert queries and takeovers**. The 18 development rollouts sum to **673 simulated seconds**, individual runs **16–66 seconds**. The model retains the common frozen ScaleBFM feedback stack across the task suite.

The accepted checkpoint exhibits completed executions across the fixed task suite using one shared reference model.

Table S7: Eighteen simulated tasks under the same shared reference checkpoint.

Task	Time (s)	Dev.	Repeat
Bottle pickup	60	yes	yes
Light switch	48	yes	yes
Drawer	66	yes	yes
Stapler	31	yes	yes
Standing fan	50	yes	yes
Faucet	28	yes	yes
Vise clamping	48	yes	yes
Drill	42	yes	yes
Hairdryer	36	yes	yes
Scissors	24	yes	yes
Wire cutting	16	yes	yes
Bakery tongs	34	yes	yes
Mango peeling	30	yes	yes
Caliper	24	yes	yes
Key	26	yes	yes
Tweezers	26	yes	yes
Umbrella	34	yes	yes
Saw	50	yes	yes

**Figure S6:** Eighteen simulated tasks executed under one accepted shared reference-policy checkpoint.

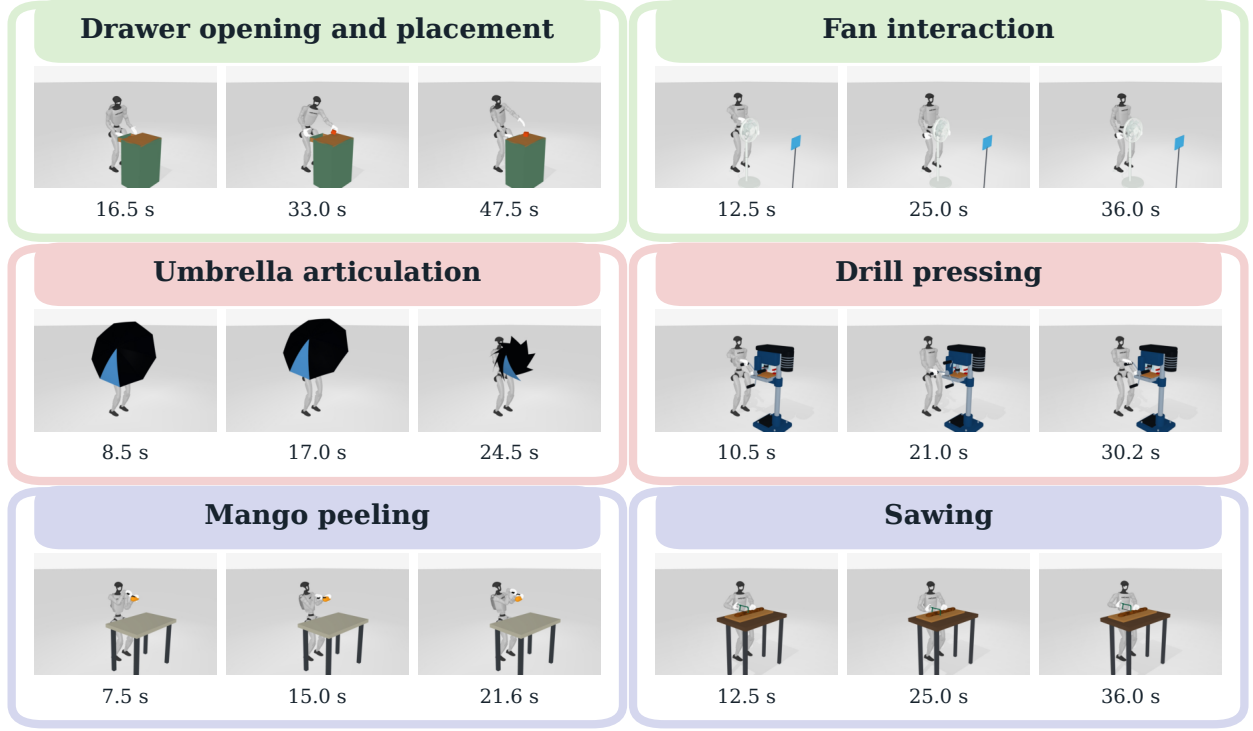


Figure S7: Six time-ordered task sequences under the same shared reference checkpoint.

S7. Physical demo registry and reproducibility-facing materials

The physical G1 archive contains six recordings across five task categories, including bottle disposal, door and cart passage, chair sit-to-stand, wall-switch operation, and standing mouse drawing. Their durations and demonstrated sequences are summarized in Table S8.

Table S8: Six recorded G1 physical demonstrations spanning five task categories.

Task	Duration	Physical sequence
Bottle disposal (real05)	101.7543 s	Grasp, turn, carry, bin release, retreat.
Door and cart (real04)	98.41 s	Door opening and pushing a cart through doorway.
Chair without hoist (real01)	24.74 s	Sit, hold and return to stand.
Chair with hoist (real02)	24.99 s	Sit, hold and return to stand.
Light switch (real03)	20.48 s	Wall-switch actuation.
Standing mouse (real06)	28.26 s	Desktop drawing with physical mouse.

References

- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 4
- Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. 4