

MOTIO: Agentic Skill Engineering for Reality-Grounded Humanoid Loco-Manipulation

SpatialtemporalAI

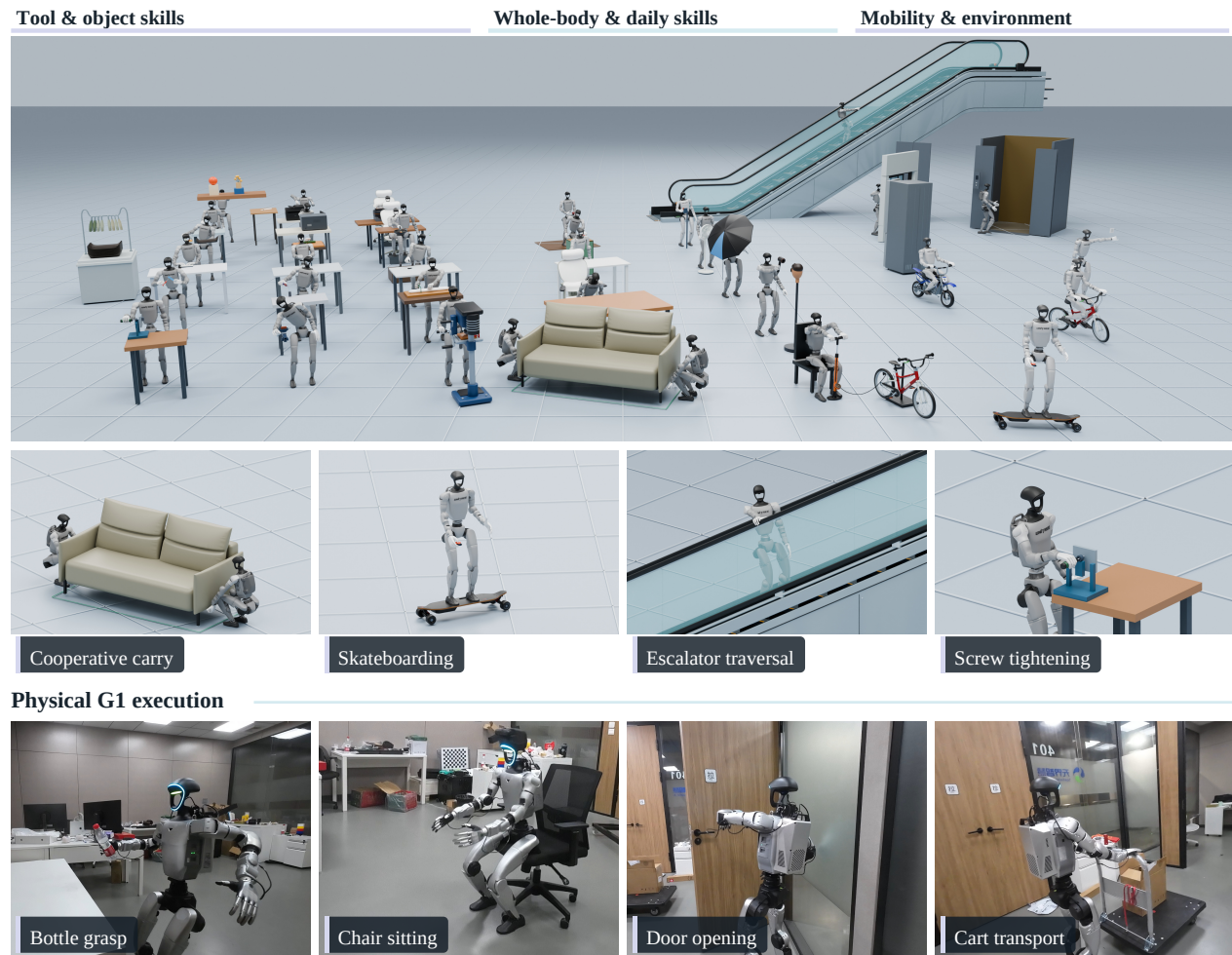


Figure 1: MOTIO develops diverse contact-rich humanoid loco-manipulation skills. Top: 36 recorded, human-reviewed skill states arranged in a shared virtual gallery, with tool and object skills (left), whole-body and daily skills (center), and mobility and environment interaction (right). Middle: cooperative carry, skateboarding, escalator traversal, and screw tightening. Bottom: physical G1 execution of bottle grasp, chair sitting, door opening, and cart transport.

Abstract

We present MOTIO, an agentic skill-engineering system that turns ordinary human-action videos, reality-grounded 3D assets, motion references, and reusable robot skills into executable humanoid loco-manipulation programs. Through in-context task and contact reasoning, an agent constructs staged robot-native interactions and improves them using physics execution, multi-frame feedback, and human validation—without task-specific teleoperation or robot demonstration collection. Agentic 3D supplies metric product and environment assets, while existing whole-body, inverse-kinematics, and dexterous-hand interfaces realize the programs. Across 33 timestamped first approvals, the median recorded agent-turn time is 97.0 minutes and request-to-approval duration is 2.80 hours. Simulation-side layout and object variation prepare a long-horizon bottle-disposal skill for a continuous 101.75-second physical G1 execution in a scanned office; physical refinement reduces screw-tightening thread residual by 76.0%. Agent-generated skills additionally supervise BC/Dagger distillation into one shared reference policy, completing 18 simulated tasks under the same checkpoint without inference-time expert intervention.

1. Introduction

Humanoid foundation controllers increasingly provide reusable walking, balancing, and whole-body tracking. Turning those capabilities into a new contact-rich task in an actual room remains an engineering challenge: the task must be grounded in object geometry, decomposed into controllable interactions, and verified under robot dynamics. Human demonstrations provide an accessible description of *what* contact events occur and *when*: reaching, grasping, turning a mechanism, regrasping, releasing, and checking the terminal object state. Such videos are naturally available, but they do not specify robot joint targets or contact-valid programs.

We investigate a different task-acquisition interface. An Agentic AI harness receives a human-action video along with the natural-language objective, metric scene or product assets, reusable motion references, robot-specific skill instructions, and successful historical programs. By reasoning over these materials in context, it constructs a staged, editable skill in robot coordinates. It can execute this program with existing controllers, inspect physical state trajectories and multi-frame visual feedback, and revise the relevant code or reference. The agent produces a physically executable, human-reviewed program that remains available for the next task.

MOTIO combines two stages. **Agentic 3D** creates reviewed metric product and scene assets, using Blender for most authored geometry and FreeCAD for parameterized, drawing- or standards-oriented models. **Agentic Skill Engineering** converts human action semantics into robot-native contacts, goals and phase transitions; MuJoCo evaluation (Todorov et al., 2012) provides the motion/contact evidence used to edit the task program. On physical G1, measured room layout, RGB-D observations, LiDAR localization, and the existing whole-body/hand execution stack support transfer of selected skills. Accepted simulator executions can also serve as expert references for a shared BC/Dagger student without introducing task-specific reward-driven reinforcement learning.

The system is grounded in actual development histories. In a HEUER bench-vice case, the agent inspected human-action video montages and prior robot programs, then iterated a 48-second right-hand grasp–turn–regrasp–clamp sequence. In a Rapesco stapler task, user feedback about palm pressing and withdrawal drove changes to explicit contact and retreat phases before a successful task event. At the system level, we analyze 33 timestamped first approvals and a diverse repertoire covering locomotion, whole-body contact, articulated mechanisms, tools, materials, and multiple humanoids. In a scanned office reconstructed with 3D Gaussian splatting (Kerbl et al., 2023), MOTIO executes long-horizon bottle disposal and door/cart interactions on a real G1, while a downstream reference policy runs 18 fixed simulated tasks using one checkpoint.

Our contributions are threefold. **(1)** We establish demonstration-conditioned *in-context skill engineering* that synthesizes executable robot-native loco-manipulation programs using existing embodied controllers and experience, without new task-specific robot demonstration collection. **(2)** We introduce a reality-grounded asset-to-skill workflow in which physical feedback and human validation drive revisions to contact stages, references, and control code. **(3)** We demonstrate this workflow across diverse accepted simulations, quantitative task-

development histories, long-horizon Real2Sim2Real executions, and downstream distillation of agent-produced expertise into a shared multitask reference model.

2. Related Work

Reusable humanoid motion and world-grounded generation. SONIC (Luo et al., 2026), ScaleBFM (Zeng et al., 2026), and Kimodo (Rempe et al., 2026) provide increasingly powerful motion tracking or generation capabilities. GRAIL (Xie et al., 2026) generates world-conditioned human–object interactions, reconstructs metric 4D interaction, retargets to humanoids, and trains task-general trackers. Object manipulation uses an object-aware adaptor with a frozen motion controller; scene interactions use adapted scene-aware tracking. It further trains egocentric visual policies for robot deployment. MOTIO uses reusable motion controllers within another task-acquisition route: human-action video supplies phase and contact semantics to a coding agent, which constructs and physically refines an editable program in robot space.

Agentic scene and program construction. RoboSimStudio (Chen et al., 2026), Real2Gym (Ren et al., 2026), Agentic RSR (Li et al., 2026), and PhysEvo (Tian et al., 2026) highlight editable interactive scenes, executable task tools, and feedback-directed improvement. USDCraft (Zhang et al., 2026a) writes and revises executable programs from partial metric geometry, rechecking geometry while authoring articulated USD simulation assets. MOTIO connects these scene-authoring ideas to long-horizon whole-body loco-manipulation through explicit staged contact programs and shared embodiment interfaces.

Human-action learning and retargeting. GPT-Policy (Cheng et al., 2026) and RoboICL (Liu et al., 2026) motivate using visual task context and prior experience to guide embodied execution. A GPT-6-Astra study (SenseNova et al., 2026) examines demonstration-derived in-context packages and harness designs on embodied benchmarks, alongside zero-shot deployment across eight real-robot configurations. Contact-aware methods such as HOI-Retarget (Shin et al., 2026), VICAR (Kalaria et al., 2026), OmniRetarget (Yang et al., 2025), and InterMimicGen (Zhang et al., 2026b) reconstruct or retarget detailed human–object motion. MOTIO instead treats video as a source of task and contact *meaning* for robot-native program synthesis. The resulting program can be tested directly in physics and reused as an expert data source.

3. Agentic Skill Engineering

Figure 2 uses the black office chair as a consistent example of the workflow: a reviewed 3D asset, an executable G1 sit–hold–stand trajectory in MuJoCo, and a human-reviewed physical G1 execution.

3.1. Reality-grounded assets and the executable task artifact

Let (g) be a natural-language task objective, (V_h) a human-action demonstration, (A) an accepted geometric asset, (E) an interactive physical environment, and (B) the robot embodiment and available low-level controllers. In Stage I, an Agentic 3D workflow searches photographs, dimensions and product documentation, creates an editable model, and delivers it to human asset review. The current catalogue groups 78 local asset families and 104 simulation-task asset families, with 59 shared families and 123 entries in their union. These are asset-family units, spanning products, support objects and scenes.

A reviewed visual asset is made task-executable by defining contact and collision geometry, articulated bodies and joints, and task-event conditions. For example, the accepted $159 \times 41 \times 53$ mm Rapesco stapler asset is linked to separate lever, magazine, anvil and press-cap contact representations. The robot program addresses a pressing target in the stapler frame and paper-stabilization targets in the paper frame; the task event checks lever motion and sustained palm/driver/paper contact. The HEUER vise similarly couples a sliding jaw and rotating screw so that a reachable palm arc can cause and measure clamping. This compact interface connects the reality-grounded model to physics and robot control; complete modeling/constraint definitions are in Supplementary Section S1.

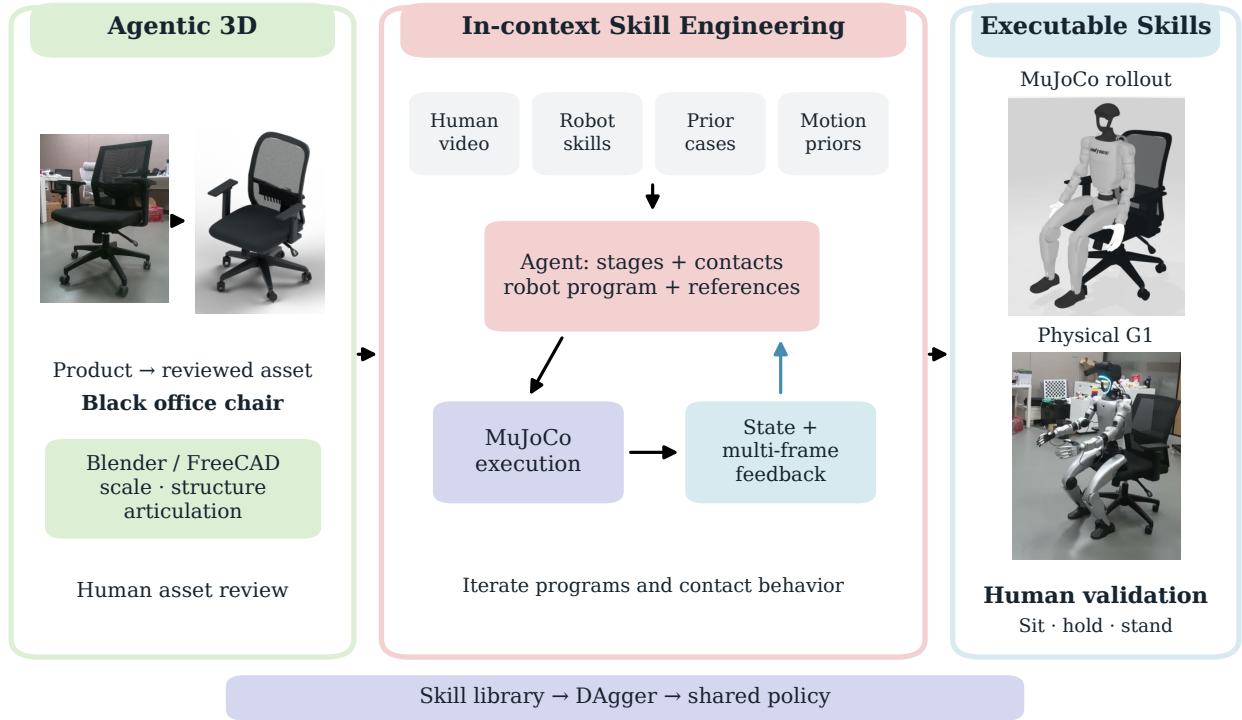


Figure 2: MOTIO workflow illustrated with G1 sitting on a black office chair. A reviewed asset enters video- and skill-conditioned program synthesis, followed by MuJoCo execution, multi-frame feedback, and human validation with physical G1 execution. Skill-library executions support downstream shared-policy learning.

Agentic 3D: reviewed assets for executable interactions

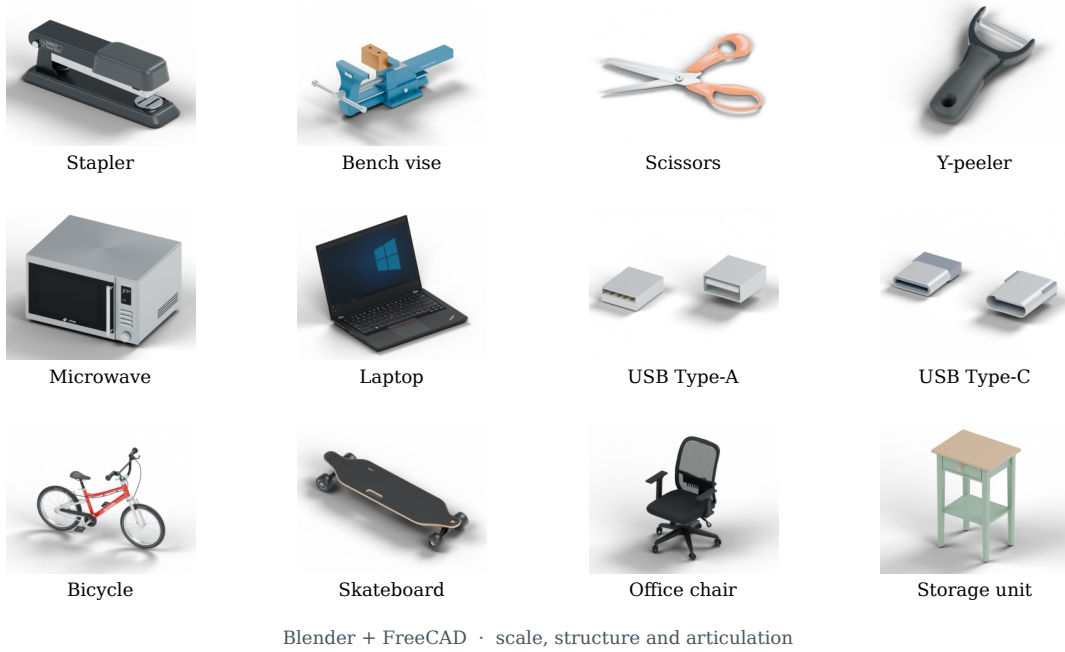


Figure 3: Representative reviewed Agentic 3D assets from everyday tools, engineered interfaces and office objects. Metric geometry and component structure are prepared for task-specific physical interaction models (Supplementary Section S1).

3.2. Video-conditioned program synthesis

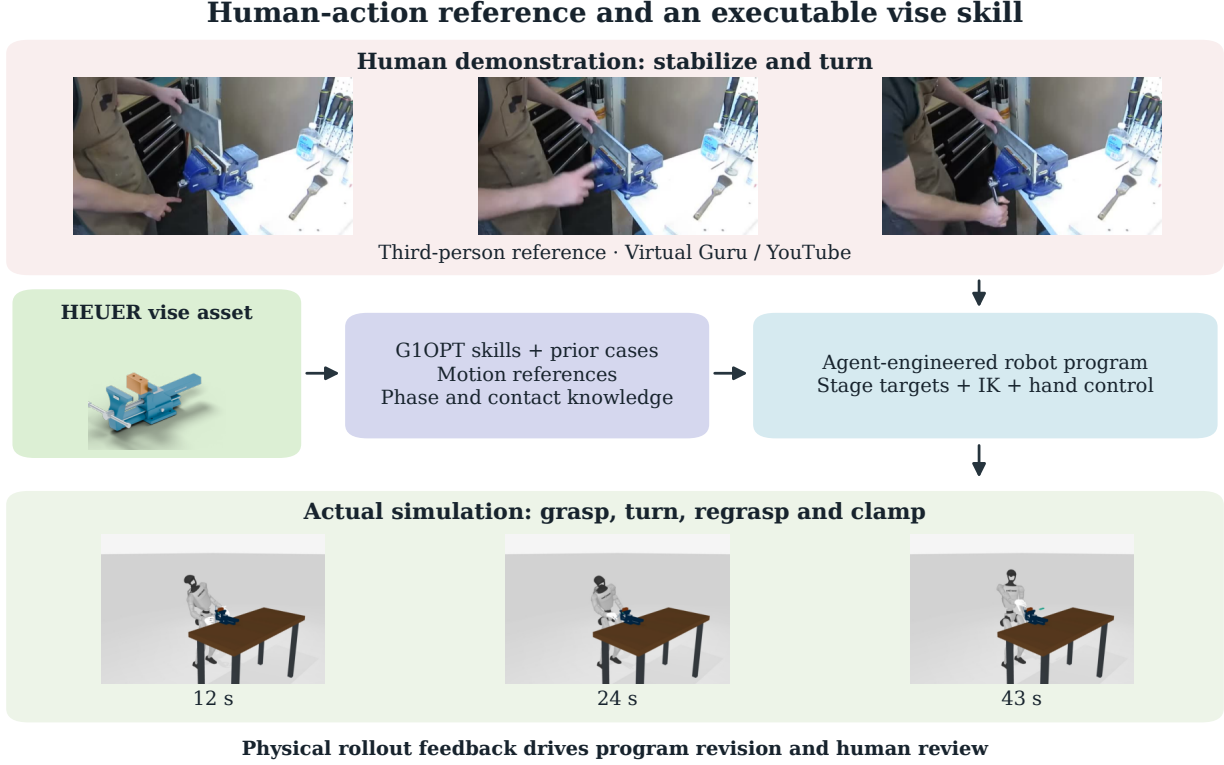


Figure 4: Video-conditioned task engineering for a HEUER vise. Top: an illustrative third-person human-operation reference (Virtual Guru / YouTube); actual agent-view video sources are detailed in Supplementary Section S2. Middle: asset and robot-skill context. Bottom: accepted G1 simulation at 12, 24 and 43 s, culminating in handle release and workpiece retention.

The task agent assembles human-action frames, task objective, accepted asset, reference motions, G1/Revo2 skill instructions, past executable cases and embodiment constraints into a single working context. The video supplies the action sequence and contact intent; the accepted asset and robot model supply scale, frames, reachable targets and physically meaningful event thresholds. The output is a phased skill program with body and hand references, control parameters and transitions based on simulated state.

The HEUER bench-vise task makes this process concrete. The original agent inspected human-action video references and retrieved robot skills and prior examples. It mapped stabilization to initially supporting the workpiece, rotation to a right-hand palm arc around the handle, regrasping to unflex/retreat/reposition/regrasp stages, and release to hand withdrawal followed by a retention test. The final accepted 48-second skill executes four seven-second half-turn cycles, producing 2.052 handle turns and reducing jaw opening from 38.000 to 29.987 mm; the workpiece remains retained after hand and support withdrawal. Figure 4 pairs the actual G1 MuJoCo trajectory with an illustrative third-person human demonstration visual; the historically inspected references are separately documented in Supplement S2.

3.3. Physics-grounded revision and human approval

For each candidate program (S_k), the agent executes a MuJoCo rollout, inspects robot/object state measurements and a multi-frame visual contact sheet, then edits targets, phase timings, contact logic or control references before the next rollout. Human feedback contributes task intent and motion-quality requirements, while the simulator reports measurable state transitions.

The Rapesco stapler records one such revision. Following a request that a press act through the palm rather

than fingers, with more natural withdrawal, the program was changed to staged palm press, unload, paper release, elbow retraction and arm lowering. The next accepted 31-second execution registered a contact-gated staple event at 14.8125 seconds; the paper remained joined while the hands withdrew and the mechanism returned. The user approved that version. The same feedback mechanism is measurable in the screw-tightening task, where successive candidates reduce maximum thread residual from 269.44 to 64.58 μm . Complete schedules and diagnostic event definitions are in Supplement S3.

3.4. Embodied deployment and optional policy reuse

For standing and mobile tasks, frozen ScaleBFM whole-body tracking executes body references with task-specific inverse kinematics, bounded hand/contact feedback and Revo2 finger schedules. MuJoCo is the principal execution and validation engine, with Isaac Sim/PhysX supporting the deformable fruit-peel interaction in the accepted mango task. To deploy the bottle task, the system uses room geometry, Livox MID-360 localization and RGB-D/SAM3 scene observations; Thor replans the staged references, validates them in MuJoCo and loads body/hand schedules to the running G1 controller.

Accepted programs and their simulated executions also supply expert state/reference pairs for one task-conditioned BC/Dagger student. The student predicts six 59-dimensional reference frames at 20-ms spacing; frozen ScaleBFM and existing contact/IK execution components realize these targets. Training relabels student-visited states from the accepted program and uses public task-event gates for recovery. The final policy uses current task/phase state and explicit geometry memory, not reference-history input. The full observation schema, target channel breakdown and training conditions are in Supplementary Section S6.

4. Reality-Grounded Loco-Manipulation

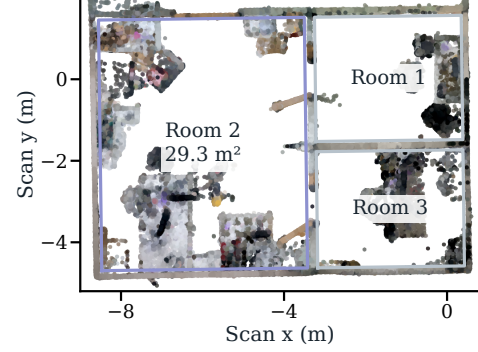
4.1. From a scanned office to an interactive task

The reconstructed office consists of three connected rooms. The bottle-disposal task uses the largest room, Room 2 (29.3 m^2). Reality-grounded representations include a 3DGS appearance model, metric point cloud and task-specific collision/navigation scene. They are aligned to the task frame used for object measurements and robot localization. Figure 5 summarizes the scanned office, interactive MuJoCo task scene, and the 31 executed layout configurations used to test transfer robustness in the same workspace.

The bottle task consists of grasp, turn, loaded locomotion, bin approach, release and retreat. It is prepared in simulation through complementary checks. **Thirty-one photo-corrected layout configurations** complete the full grasp–carry–release task with continuous carrying grasp (61.2–138.6 seconds). **Forty-four distinct bottle-property settings** span diameter 58–68 mm, height 220–235 mm and mass 20–600 g and each pass the opposing-finger **hand-fixture** grasp window at $8 < t < 10$ seconds, followed by release. On the same twelve matched settings, contact refinement improves full-window grasp from 11/12 to 12/12. **Three perception-conditioned whole-body variants** complete eight-stage sequences with maintained carrying grasp, release and return to stand. Supplementary Section S4 and Figure S2 retain the complete layout and object-parameter coverage, while Table 1 summarizes the preparation and deployment outcomes.

For physical execution, the observed layout and bin measurement determine a staged trajectory that is first checked in MuJoCo. The final continuous real05 G1 episode executes bottle pickup, turning, loaded walking, bin release and retreat in **101.7543 seconds**, without execution-time retry or task-specific robot demonstration collection. A separate **98.41-second** physical recording demonstrates door opening followed by cart transport. Figure 6 shows selected frames from both continuous physical episodes, with corresponding insets highlighting contact and task details. Chair sit-to-stand, switch pressing and mouse drawing provide further physical examples, while broader simulation tasks span moving support, balancing, articulated mechanisms, multi-robot carrying, tools, materials and digital interfaces. Supplement S7 contains the full six-video register; this main section remains centered on the complete Real2Sim2Real route.

a Office appearance · 3D Gaussian splats

b Metric scene geometry · point cloud

c Interactive MuJoCo task scene

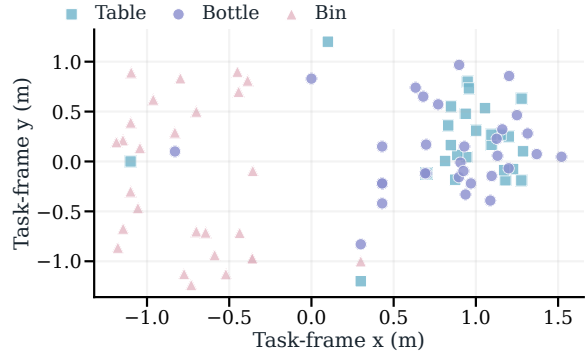
d Layout adaptation · 31 configurations


Figure 5: Reality-grounded office reconstruction and task preparation. (a) Room 2 appearance rendered from the scanned 3D Gaussian model. (b) Metric scan geometry across three connected rooms; Room 2 spans 29.3 m^2 . (c) The interactive MuJoCo bottle-to-bin scene. (d) Table, bottle and bin locations across 31 executed layout configurations.

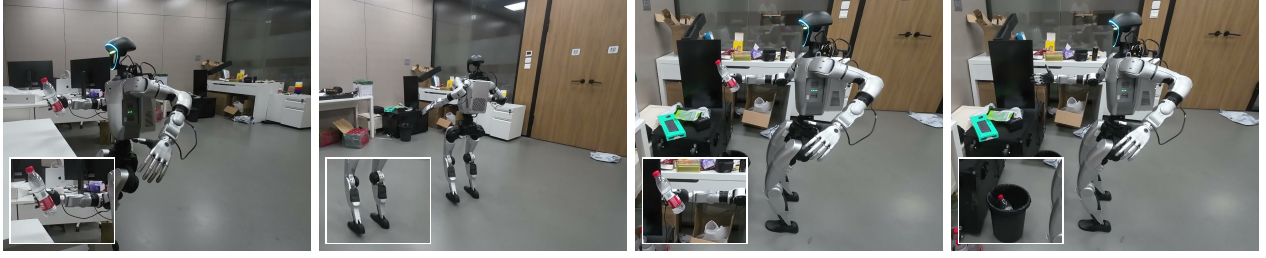
Table 1: Simulation preparation and physical G1 execution.

Study	Variation or sequence	Observed result
<i>Simulation preparation</i>		
Full-task layouts	31 configurations	31/31 complete sequences
Bottle properties	44 unique settings	44/44 full-window grasps
Paired grasp refinement	Same 12 settings	11/12 → 12/12 full-window retention
Perception-conditioned execution	3 whole-body variants	3 complete eight-stage sequences
<i>Physical G1 execution</i>		
Bottle pickup, carry and disposal	Continuous episode	101.75 s
Door opening and cart transport	Continuous episode	98.41 s

Grasp-window retention is measured over 8–10 s; each whole-body variant executes all eight stages.

Bottle grasp, carry and disposal

101.75 s



Grasp

34 s Carry

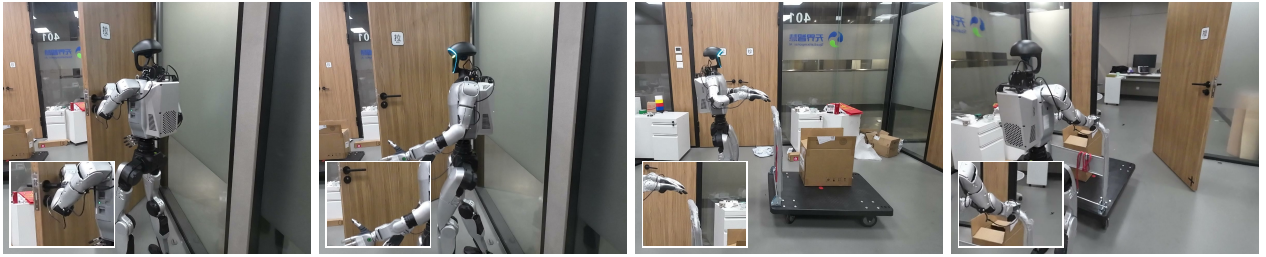
69 s Position

97 s Dispose

100 s

Door opening and cart transport

98.41 s



Grasp handle

23 s Open door

31 s Reach cart

68 s Transport

87 s

Figure 6: Physical G1 deployment. Top: selected moments from the continuous 101.75-second bottle grasp, carry and disposal episode. Bottom: door opening followed by cart transport in a 98.41-second episode. Insets magnify contact and task details from the corresponding frames.

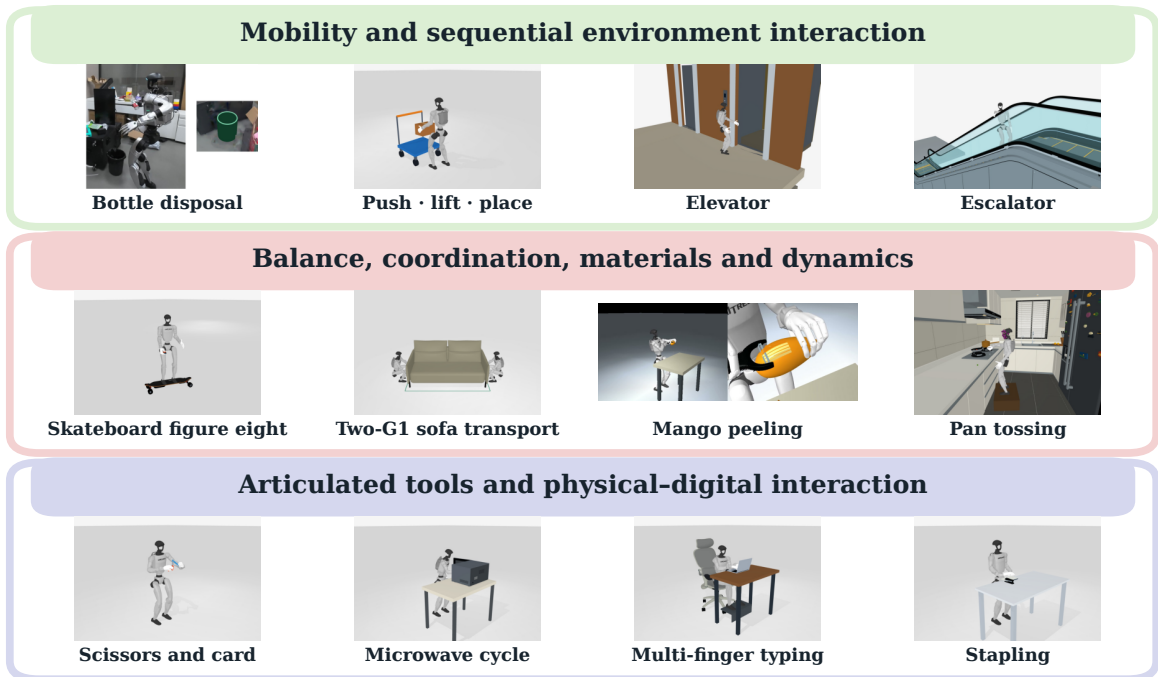


Figure 7: Representative accepted simulation tasks spanning long-horizon mobility, sequential interaction, balance, multi-robot cooperation, articulated tools, deformable materials and physical–digital interfaces.

5. Quantitative Skill Development

5.1. Development time and execution-feedback budget

How much development does a skill require?

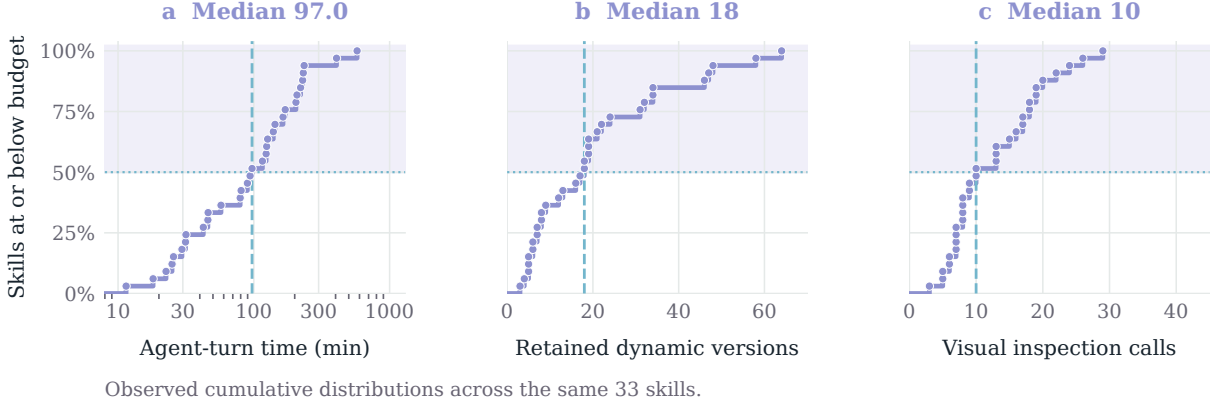


Figure 8: Development budget across 33 skills with matched first human approval: empirical cumulative distributions for agent-turn time, retained dynamic trajectories and visual inspection calls. Dashed vertical lines mark medians.

We analyze 34 consecutively registered skill families (Cases 40–73), 33 of which have precisely matched task requests and first human-approval timestamps. Recorded **agent-turn time** includes tool and resource waits within open turns; **request-to-approval time** is the calendar interval including periods between turns and human review. The median first-approval budget is **97.0 agent-turn minutes** (IQR 42.5–170.3) and **2.80 calendar hours** (IQR 1.33–9.80). Median first-approval development includes **two agent turns**, **18 retained dynamic versions**, **ten visual inspections** and **one substantive human revision prompt**.

An empirical budget summary provides the scale of the engineering process: **25/33 (75.8%)** reach first approval within at most three hours of recorded agent-turn time; **21/33 (63.6%)** within 20 retained dynamic versions; **17/33 (51.5%)** within ten visual-inspection calls. Figure 8 reports empirical cumulative distributions for the three budget measures across the 33 timestamped skills. Full per-skill time profiles and three histories of expanding human objectives (two-G1 sofa carrying, skateboard and flexible bag) move to Supplementary Section S5. These illustrate how an accepted first skill can be modified for a richer goal.

Table 2: Development budget to first human approval across 33 skills (median and interquartile range).

Measure	Median	IQR
Agent-turn (min)	97.0	42.5–170.3
Request-to-approval (h)	2.80	1.33–9.80
Agent turns	2	1–2
Retained dynamic versions	18	7–31
Visual inspections	10	7–18
Human revision prompts	1	0–1

5.2. Measurable physical refinement

Physical state feedback changes task outcomes, not merely the appearance of trajectories. In the screw-tightening case, maximum thread residual falls from **269.44 to 64.58 μm (76.0% reduction)** across development, and

Physical feedback improves screw-tightening accuracy

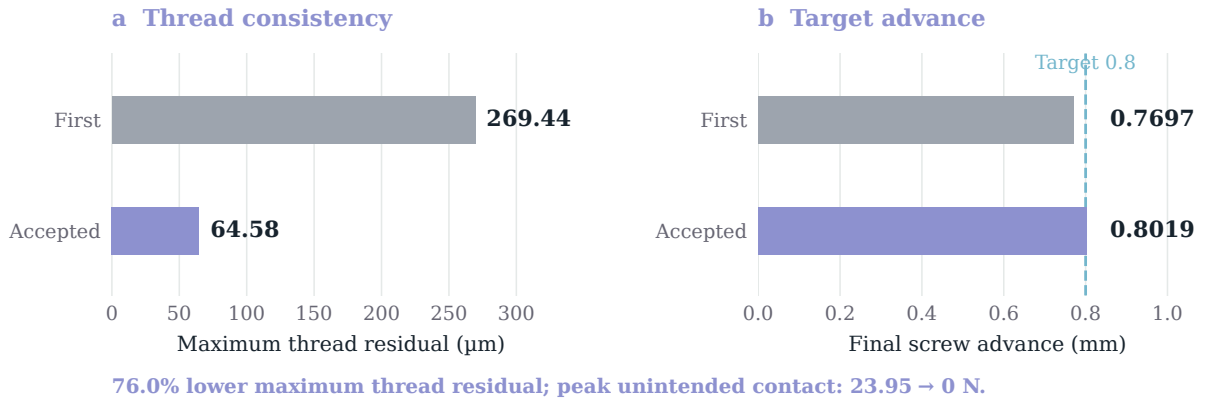
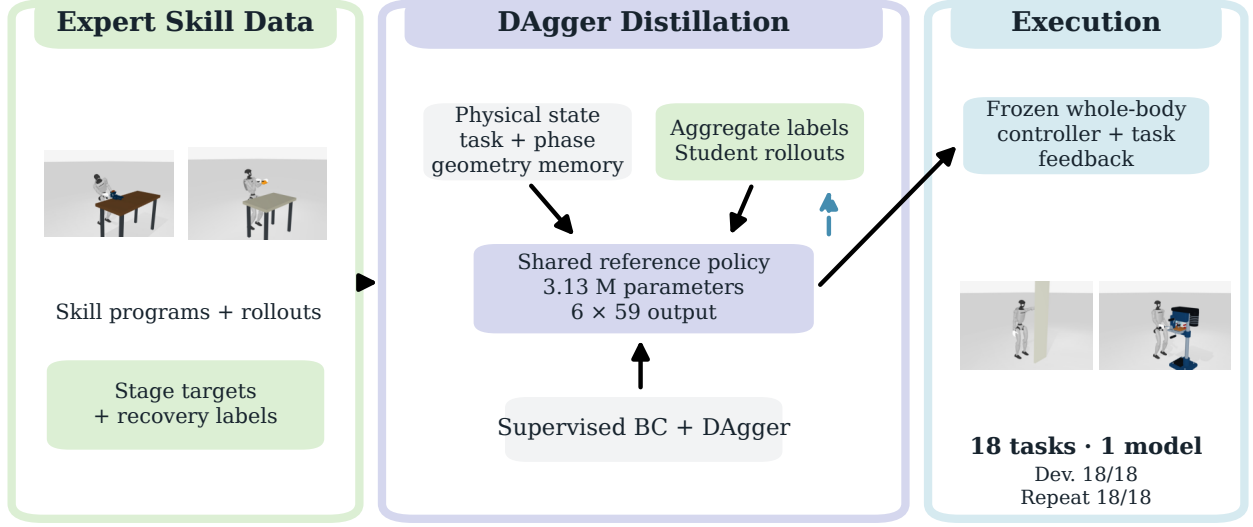


Figure 9: Physical refinement of screw tightening from first to human-approved final candidate: maximum thread residual 269.44 to 64.58 μm (76.0% lower) and final advancement 0.7697 to 0.8019 mm against 0.8 mm target.

screw advance moves from **0.7697 to 0.8019 mm** for a **0.8 mm** target. Recorded peak unintended contact falls from 23.95 N to zero. The compact main quantitative figure directly compares first and accepted candidates; Supplementary Section S5 retains the complete full01–full04 candidate sequence and its physical outcome data.

6. One Shared Multitask Reference Policy

MOTIO’s accepted skill programs can be executed directly or aggregated as expert references for a task-conditioned student. Behavior cloning (BC) initializes reference generation; dataset aggregation (Dagger) (Ross et al., 2011) labels visited student states using each agent-produced expert. The shared student has **3,127,287 parameters** and produces **6 $\times$ 59** references, consumed by the frozen ScaleBFM and existing bounded task-feedback execution stack. The ten-task extension dataset contains **63 episodes and 92,130 frames**, including 77,130 student-execution frames and 142 recovery-flag frames counted within the total. Across joint system-development stages, completion on ten newly added tasks rises from **2/10 at BC** to **5/10 (D1)**, **7/10 (D2)**, **8/10 (D3 and D4)**, **9/10 (D6)** and **10/10 (D7)**. The final accepted **single checkpoint** completes **18/18 fixed simulated tasks** in development and **18/18 same-condition repeated runs**, with **zero inference-time expert queries or takeovers**. The eighteen development rollouts sum to 673 simulated seconds (16–66 seconds each). This outcome demonstrates reuse of agent-engineered expert behavior within one reference model, without task-specific RL reward design. The downstream supervision pipeline and development results are shown in Figures 10 and 11; detailed I/O, trained-task durations, and the full execution gallery appear in Supplementary Section S6.



Agent-engineered executions → Dagger → shared multitask references

Figure 10: Agent-engineered simulated skill executions provide BC initialization and Dagger expert references for a shared reference generator, executed through the frozen ScaleBFM whole-body controller and retained task-feedback interfaces.

Downstream shared-policy development

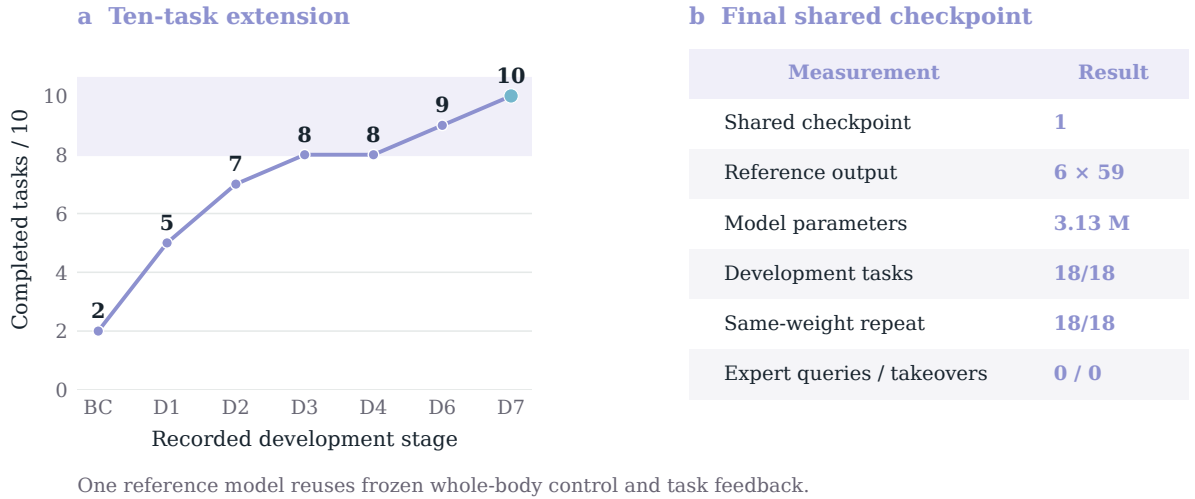


Figure 11: Shared-policy development across ten added tasks (BC 2/10 to D7 10/10) and final one-checkpoint eighteen-task performance (18/18 development and 18/18 same-condition repeat; zero inference-time expert queries/takeovers). The stage trace includes program, representation and execution-interface refinements.

7. Conclusion and Outlook

MOTIO turns human-action context, reality-grounded assets and existing robot capabilities into executable, contact-rich humanoid loco-manipulation programs. The agent engineers tasks in robot space, learns from physical execution and human review, and produces skills that transfer to selected real G1 tasks and supply shared-policy training data. Verified case studies, first-approval budgets, contact-accuracy refinement and one-checkpoint policy execution establish the present scope. Parallel agent development, unified validation and accumulated execution data offer a path toward substantially larger reusable skill libraries.

References

- Kaiyuan Chen, Shuangyu Xie, Uksang Yoo, Xinyue Zhu, Cael Magner, Alvin Zhu, Daehwa Kim, Eric Heiden, Erwin Coumans, Miles Macklin, Spencer Huang, and Ken Goldberg. RoboSimStudio: An interactive toolkit for human and agent rapid real2sim of robot manipulation tasks, 2026. URL <https://arxiv.org/abs/2610.18835v2>. Version 2, revised October 14, 2026; arXiv abstract title. 3
- Dongzhou Cheng, Taoran Yi, Ye Fang, Xingwu Zhang, Fan Feng, Yixuan Li, Gengxiong Zhuang, Rongze Wang, Shuai Yang, Wei Song, et al. In-context robot learning with VLM agents, 2026. URL <https://arxiv.org/abs/2609.19138>. 3
- Dvij Kalaria, Bike Zhang, and S. Shankar Sastry. VICAR: Video imitation with contact-aware retargeting for humanoid interactive skills, 2026. URL <https://arxiv.org/abs/2610.13079>. 3
- 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. 2
- Yihan Li, Yating Feng, Shengjiu Sun, Jianing Chen, Hao Ren, Bowen Yang, Weisheng Xu, Qiwei Wu, Hui Cheng, and Renjing Xu. Agentic RSR: Real-to-sim-to-real through scene reconstruction and execution-grounded robot policies, 2026. URL <https://arxiv.org/abs/2610.10479v2>. 3
- Fangcheng Liu, Yeqing Shen, Anda Cheng, Weishi Mi, Chao Tang, Chenyuan Liu, Yushun Xiang, Tingguang Li, Yong-Lu Li, and Yehui Tang. RoboICL: Embodied in-context learning with GPT-6 Astra, 2026. URL <https://arxiv.org/abs/2609.34261>. 3
- Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eae4592, 2026. doi: 10.1126/scirobotics.aed4592. URL <https://arxiv.org/abs/2511.07820>. 3
- Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, et al. Kimodo: Scaling controllable human motion generation, 2026. URL <https://arxiv.org/abs/2603.15546>. 3
- Kerui Ren, Yingxiang Xu, Kaiwen Song, Linning Xu, Bo Dai, Mulin Yu, and Tao Lu. Real2Gym: Building gyms from videos, bringing skills to robots, 2026. URL <https://arxiv.org/abs/2609.37089>. 3
- Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635. PMLR, 2011. URL <https://proceedings.mlr.press/v15/ross11a.html>. 10
- SenseNova, Joint Research Team, Zhongang Cai, Ziang Cao, Tailai Chen, Han Chong Chua, Linjun Dai, Hanming Deng, Hokin Deng, Haiwen Diao, Chenjing Ding, Weichen Fan, Xiao Fu, Zerui Gong, Chenyang Gu, Zexin He, Zhi Hou, Bo Li, Qinyue Li, Quanzhou Li, Xiaowan Li, Zhichao Liao, Dahua Lin, Zilong Lin, Bo Liu, Chunxiao Liu, Pengfei Liu, Tianqi Liu, Ziwei Liu, Chen Change Loy, Lewei Lu, Jiapeng Luo, Yuxiao Ma, Liang Pan, Xingang Pan, Oscar Qian, Ziyi Shang, Yuyao Shao, Dawei Su, Dacheng Tao, Qingyi Tao, Quan Wang, Ruisi Wang, Xiaogang Wang, Ziyi Wang, Beichen Wen, Donghao Wu, Jiaxu Wu, Silei Wu, Xiaoming Wu, Zhonghua Wu, Hongcan Xiao, Weiye Xu, Atsushi Yamashita, Lei Yang, Zhe Yang, Baiqiao Yin, Wanqi Yin, Hanfei Zhao, Tianbo Zhong, Long Zhuo, and Jing Zuo. Has gpt-6-astra achieved embodied intelligence?, 2026. URL <https://arxiv.org/abs/2610.22471>. 3
- Jihwan Shin, Adrià López Escoriza, Junzhe He, Matthias Heyrman, and Marco Hutter. HOI-Retarget: Contact-centric retargeting for human-object interaction, 2026. URL <https://arxiv.org/abs/2609.34674>. 3
- Wenqing Tian, Zeyu Zhang, Zhaocheng Liu, Fengwei Liu, Qiang Liu, and Liang Wang. PhysEvo: Astra can act, let it, 2026. URL <https://arxiv.org/abs/2610.08995>. 3
- 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. 2
- Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from world models, 2026. URL <https://arxiv.org/abs/2606.05160v2>. Version 2, revised October 19, 2026. 3
- Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C. Karen Liu, Rocky Duan, and Guanya Shi. OmniRetarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction, 2025. URL <https://arxiv.org/abs/2509.26633v3>. arXiv version 3, revised June 15, 2026. 3
- Weishuai Zeng, Kangning Yin, Xiaojie Niu, Shunlin Lu, Weixiang Zhong, Jiahe Chen, Feiyu Jia, Xiao Chen, Zirui Wang, Furui Xu, et al. Scaling behavior foundation model for humanoid robots, 2026. URL <https://arxiv.org/abs/2607.15163>. 3
- Chuanrui Zhang, Zaijia Yang, Duomin Wang, Lu Shi, Daquan Zhou, Ruihua Zhang, and Ziwei Wang. Usdcraft: Geometrically grounded

programmatic modeling of articulated 3d assets for simulation. *arXiv preprint arXiv:2610.11322*, 2026a. doi: 10.48550/arXiv.2610.11322. URL <https://arxiv.org/abs/2610.11322v1>. 3

Yucheng Zhang, Sirui Xu, Jinhong Li, Liuyu Bian, Anatulya Nandi, Derek Zhang, Xiangchen Liu, Xueting Li, Umar Iqbal, Yu-Xiong Wang, and Liang-Yan Gui. InterMimicGen: Scaling humanoid loco-manipulation through self-evolving motion imitation, 2026b. URL <https://arxiv.org/abs/2610.06850>. 3