Reins: Frontier Multimodal Models Driving Real Robot Arms on Long-Horizon Tasks, Zero-Shot

Pengfei Ye[1] **Zhiyang Dou**[1] **Yuxiang Ma**[1]
Haonan Chen[2] **Heng Yang**[2] **Edward H. Adelson**[1]
[1]MIT CSAIL [2]Harvard University pfy712@mit.edu

Abstract: Frontier models such as GPT-6.1 Sol, GPT-6 Astra and Claude Opus 5.5 can now plan and call tools, yet they still struggle to drive real robots through long-horizon, contact-rich manipulation, blind while the arm moves and while they think. We present Reins, a harness in which an off-the-shelf model operates a dual-arm robot with 3D perception tools and atomic actions that report contact and grip. The model chains actions in one call, each with a condition that must hold, stopping at the first that fails; a second instance, the narrator, watches the cameras continuously and reports what changed with every result. From images, Reins also builds a simulation-ready twin of the scene for continual policy learning. With natural-language prompts and no task-specific code, GPT-6.1 Sol succeeds in 88% of real episodes on five tasks; without the harness, it fails every shell game, taking 8 times as long. Three frontier models perform alike but differ in tracking a scene that changes while the actor thinks. Code will be released upon publication.

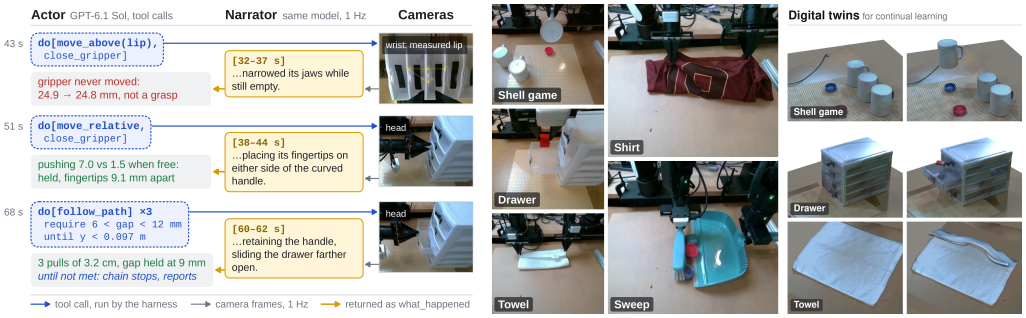


Figure 1: An overview of Reins. Left: three calls of a Sol drawer episode; results report what the robot measured, a chain of pulls stops when its condition fails, and the narrator's account arrives with the next result. Middle: the five tasks. Right: sim-ready digital twins, each built from one head-camera frame, at rest and running in MuJoCo (App. C), for continual learning; only object dynamics are shown.

1 Introduction and Related Work

Frontier models that plan and call tools [1, 2] now run as robot policies without robot training, mostly in simulation, where benchmarks find a gap between understanding and executing a task [3–8]. On real hardware, Show-Harness [9] moves a Franka and a bimanual AgileX in 2–4 cm steps, mostly pick-and-place. Trained policies struggle too: on RoboDojo's real *Sweep Blocks*, ten vision-language-action models trained on 100 demonstrations each never succeed [10].

Earlier language-model systems compose skills, write programs or spatial targets, or feed scene descriptions back to the model [11–21]; others learn control from robot data [22–28] or sequence learned skills [29–31]. In Reins, the model drives the arms through atomic actions, with no skill library or task code; two choices set it apart. First, the model attaches stopping conditions to each chain of actions, checked at every step, as in guarded moves [32, 33] and the action conditions of task and motion planning [34], but written at run time; Show-Harness runs its action chunks open-loop [9]. Second, another instance of the model, the narrator, watches every camera frame while the actor thinks and keeps a running account; VoLo's agent monitors its own execution at 0.2 Hz [35]. We contribute (i) Reins with its guarded chains, (ii) the narrator and its latency analysis, and (iii) a real-robot evaluation of three frontier models on five long-horizon tasks, with time and cost.

10th Conference on Robot Learning (CoRL 2026), Austin TX, USA.

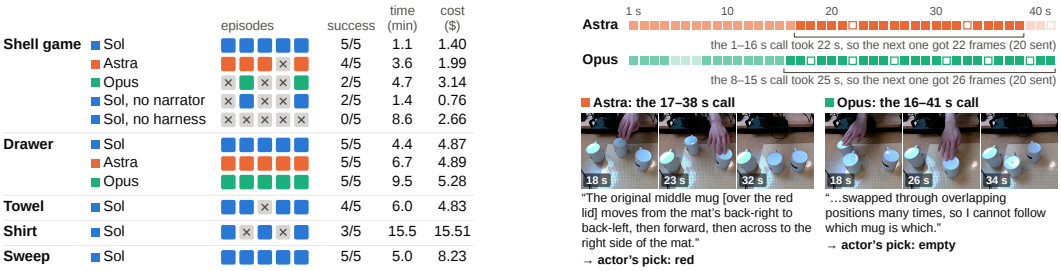


Figure 2: Results and the narrator. (a) Five episodes each (×: failure); time: median minutes; cost: mean API cost. (b) Astra's and Opus's narrators (standard tier) in a shell game: one square per frame, shaded by call.

2 Method

Actor loop. At decision k, the actor, an off-the-shelf model π given a task prompt g, outputs a tool call or chain of calls $a_k \sim \pi(\cdot \mid g, h_k)$; its result $y_k = (r_k, I_k, q_k, E_k)$ (measurements, new images, both arms' state and the narrator's entries since y_{k-1}) joins the history h_{k+1} (App. A). An episode of K decisions lasts $T = \sum_k (\lambda_k + d_k)$, model latency plus execution. A chain of thought takes time: latency is 41–78% of T in our runs, so each decision should cover several checked steps (1.6–3.4 on average), and the scene may change while the actor thinks. **Perception and atomic actions.** Of the eighteen tools (App. B), `find` (SAM 3 [36] on RGB-D) and `measure` turn an object name or image point into a 3D point in the base frame, to about 1.6 mm with the arm's own wrist camera. Atomic actions move an arm, tilt its tool or work its gripper; `move_until_contact` stops when joint current indicates contact, and `close_gripper` reports whether the fingers hold anything. **Guarded chains.** `do` executes a chain $P = \big((p_i, \rho_i, u_i, n_i)\big)_{i=1}^{n}$ of up to eight calls and sixteen runs. Step i runs tool p_i up to n_i times until u_i holds, and each result $r_{i,m}$ must pass ρ_i. The chain stops after its last step or at the first run that fails its check or uses up its repeats:

$$(i, m)^\star = \min\big\{(i, m) \;:\; \neg\rho_i(r_{i,m}) \;\vee\; \big(m = n_i \wedge \neg u_i(r_{i,m})\big)\big\}. \tag{1}$$

Narrator. A second instance ν of the model watches the head camera at 1 frame/s in one ongoing conversation. Call j receives the frames in $W_j = (\tau_{j-1}, \tau_j]$ (at most 20, plus wrist frames while an arm moves) and, after latency ℓ_j, returns an entry $e_j = \nu(W_j, e_{<j})$ that reaches the actor as `what_happened`. The next window closes when the call returns, $\tau_{j+1} = \tau_j + \ell_j$, so

$$|W_{j+1}| = \ell_j, \qquad \ell_j \;\le\; t - \tau_j \;<\; \ell_j + \ell_{j+1} \quad \text{for } t \in [\tau_{j+1}, \tau_{j+2}): \tag{2}$$

each window lasts as long as the previous call, and the newest frame reported is 1–2 latencies old.

3 Experiments and Findings

Setup and tasks. Two 6-DoF AirBot Play arms with stock parallel grippers share a table, with RGB-D cameras on the head and each wrist; episodes start at the zero pose and the operator judges success. We run five episodes per condition with GPT-6.1 Sol (Ultrafast tier), GPT-6 Astra and Claude Opus 5.5, at high reasoning effort. In the *shell game*, the first of three mugs lifted after a 20–30 s shuffle must cover a hidden red object; the *drawer* is opened by its 3–4 mm lip, filled with a block and closed; a *towel* is folded twice, a *shirt* sleeves-in and then twice; in *sweep*, one arm holds a dustpan and the other brushes objects in. The drawer and sweep prompts add grip and pose hints.

The narrator and the harness. All three models open the drawer every time; the shell game, whose scene changes while the actor waits, separates them. Without the narrator, Sol wins only 2 of 5. By Eq. (2), a faster narrator sees more: Sol's (Ultrafast tier) answers in 1–6 s and sees a shuffle in 8–11 windows; Astra's and Opus's (standard tier) take 3–40 s and see it in 2–4 (Fig. 2b). Astra's narrator still tracks the red object in 4 of 5 episodes, Opus's in none; Opus's two wins were guesses. Without the harness (raw images, straight-down moves, `wait`, `look`), Sol never lifts the correct mug: four times it closes the gripper 6–9 times in 60 calls but lifts no mug, and once it lifts the wrong one.

4 Limitations and Future Work

Limitations. Task knowledge enters only through natural-language prompts; contact is sensed only from joint current and gripper width; we test one platform, five episodes per condition; and a slow narrator reports fast changes late (Eq. 2). **Future work.** Every episode is fully logged (100 Hz state, images, text); we will turn successful episodes into reusable chains and checks, rehearsed in simulation twins (App. C), so that the same model finishes tasks faster, from shorter prompts.

References

[1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[2] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

[3] W. Zhang, K. Wang, Y. Ouyang, X. Huang, L. Li, K. Su, W. Jin, W. Chai, H. Liang, Z. Dou, et al. An unexpected robot policy: Early evaluations of GPT-6 Astra on RoboDojo and beyond. *arXiv preprint arXiv:2609.24170*, 2026.

[4] J. X. Yang, A. Khetan, and S. Zhang. Astranomy: An empirical study of GPT-6 Astra as a native robot policy. NVIDIA Seattle Robotics Lab technical report, https://research.nvidia.com/labs/srl/projects/astranomy/, October 2026.

[5] M. Espinola, Q. Zhang, and Y. Xiang. From reasoning to action: Benchmarking frontier AI models on LIBERO-10 robot manipulation tasks. *arXiv preprint arXiv:2610.30051*, 2026.

[6] D. Huang, J. Wang, J. Wu, T. Ran, L. Guo, S. Zhang, C. Men, S. Wang, Y. Zhou, M. Yao, et al. Frontier LLMs enter the world of robotics: A cross-model evaluation on RoboColiseum. *arXiv preprint arXiv:2610.27865*, 2026.

[7] Y. Zhou, B. Hou, Y. Fu, Y. Lyu, Y. Ye, S. Shan, C. Guo, J. Liu, R. He, C. Shan, et al. Large multimodal models for direct robot control: GPT-6 as a beginning. *arXiv preprint arXiv:2610.28259*, 2026.

[8] B. Y. Tsui, A. Y. Fang, and T. J. Hwu. Demonstration-free robotic control via LLM agents. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026.

[9] Y. Chen, Z. Bai, Z. Cao, W. Zeng, K. Q. Lin, Y. Lin, G. Liang, K. Y. Ma, Q. Huang, and M. Z. Shou. Show-Harness: Just a VLM agent can play robots. *arXiv preprint arXiv:2609.10522*, 2026.

[10] T. Chen, Y. Chen, Z. Li, J. Tang, K. Su, H. Lu, W. Wan, B. Chen, S. Liu, H. Yan, et al. RoboDojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026.

[11] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. Do as I can, not as I say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[12] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg. Text2Motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365, 2023.

[13] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng. Code as Policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500, 2023.

[14] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg. ProgPrompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302*, 2022.

[15] S. H. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor. ChatGPT for robotics: Design principles and model abilities. *IEEE Access*, 12:55682–55696, 2024.

[16] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. VoxPoser: Composable 3D value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[17] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. ReKep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.

[18] F. Liu, K. Fang, P. Abbeel, and S. Levine. MOKA: Open-world robotic manipulation through mark-based visual prompting. *arXiv preprint arXiv:2403.03174*, 2024.

[19] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. Inner Monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[20] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, et al. Socratic Models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[21] Z. Liu, A. Bahety, and S. Song. REFLECT: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023.

[22] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[23] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion Policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

[24] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[25] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[26] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[27] Physical Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[28] Gemini Robotics Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, et al. Gemini Robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[29] Google DeepMind. Gemini Robotics-ER. Gemini API documentation, https://ai.google.dev/gemini-api/docs/robotics-overview, 2026.

[30] R. Li, Y. Zhou, Y. Zhu, K. Chen, J. Wang, S. Wang, K. Hu, M. Yu, B. Jiang, Z. Su, J. Ma, X. He, Y. Shen, Y. Yang, G. Ren, M. Yao, W. Wang, and Y. Mu. RoboClaw: An agentic framework for scalable long-horizon robotic tasks. *arXiv preprint arXiv:2603.11558*, 2026.

[31] M. Ahn, D. Dwibedi, C. Finn, M. G. Arenas, K. Gopalakrishnan, K. Hausman, B. Ichter, A. Irpan, N. Joshi, R. Julian, et al. AutoRT: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963*, 2024.

[32] P. M. Will and D. D. Grossman. An experimental system for computer controlled mechanical assembly. *IEEE Transactions on Computers*, C-24(9):879–888, 1975.

[33] T. Lozano-Pérez, M. T. Mason, and R. H. Taylor. Automatic synthesis of fine-motion strategies for robots. *The International Journal of Robotics Research*, 3(1):3–24, 1984.

[34] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez. Integrated task and motion planning. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:265–293, 2021.

[35] S. Chen, H. Hadfield, A. Zook, M. A. Uy, C. H. Song, E. Coumans, X. Yang, F. Ladhak, Q. Qu, S. Birchfield, J. Tremblay, and V. Blukis. VoLo: A physical orchestrator for open-vocabulary long-horizon manipulation. *arXiv preprint arXiv:2606.07723*, 2026.

[36] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al. SAM 3: Segment Anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

[37] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[38] E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

[39] A. Jacobson, L. Kavan, and O. Sorkine-Hornung. Robust inside-outside segmentation using generalized winding numbers. *ACM Transactions on Graphics*, 32(4):33:1–33:12, 2013.

A Context, Caching and Memory

What the actor keeps. The text of an episode stays in the history until the end: every reasoning step, tool result and measurement. Images do not. A result's images remain available for the following decisions, but once the prompt has grown by 30k tokens since the last compaction, all images except those of the newest image-bearing turn are replaced at once by one-line stubs. What the model derived from an old image remains: the object ids that `find` registered, the pixels it measured and the results. A chain returns only the images of its last run.

Why rarely and in bulk. On the provider route we use, the prompt cache only reuses a prompt that extends one it has already served in full, so any edit of earlier history costs one full re-read. Two alternatives performed worse. Compacting whenever the text exceeded a fixed budget edited the history on every call: after the first compaction of a 65-call episode, every call re-read 45–54k tokens at full price, 90% of that episode's uncached tokens. A sliding window that kept the two newest image turns and edited the history before every call reduced the cached prefix to the 6.3k-token system prompt and raised the actor's cost per call from \$0.24 to \$0.85 on the Ultrafast tier, without making calls faster. Compaction triggered by growth costs one re-read per compaction, after which the cache holds again. For Claude, whose provider caches only a prefix that the request marks, the harness adds the marker.

The narrator's context. The narrator reports by exception: it replies only with what a window adds (an entry, a correction, or `NO CHANGE`), and the harness stamps the reply with the window's time span and appends it; consecutive quiet windows are merged into one line. An earlier version that rewrote the whole account on every call spent 11 s and 666 output tokens on a window in which nothing had changed. The calls form one continuing conversation, so each window is a new turn on a cached prefix; re-sending the account in a fresh prompt for each window was cached in none of 41 calls. The conversation restarts every 60 frames (about a minute), seeded with the episode's first frame and the account so far, so old frames are dropped in bulk with one uncached read per restart. We count frames rather than tokens because models encode a frame differently: Claude Opus 5.5 spends about 600–700 tokens on a frame that GPT-6 encodes in about 240, so a 15k-token budget restarted Opus's narrator every 1.9 calls (43 restarts in one drawer episode, with 32% of its input cached). A 60k-token ceiling still bounds every model's conversation.

Memory within an episode. Besides the text history and the narrator's account, four records persist across calls. (i) *Objects*: `find` registers each object as an id with a 3D position, drawn on every later image and listed in every result. (ii) *Measured shapes*: `measure` can turn a visible contour, such as a cloth edge or a drawer lip, into base-frame points that remain valid after the arm occludes it, and `follow_path` can travel through them. (iii) *A task ledger*: `plan` records the steps and `mark_step` the evidence for each; `done` is refused while a planned step is pending. (iv) *What each arm holds*, with the finger gap measured at the grasp. The narrator names identical objects by where it first saw them ("the original middle mug"), which lets an identity survive a shuffle.

Measured. Table 1 summarizes the 45 counted episodes.

model	input tokens from cache: actor	input tokens from cache: narrator	compactions per episode median (max)	largest prompt median (max)
GPT-6.1 Sol	90%	73%	1 (5)	61k (171k)
GPT-6 Astra	91%	68%	1 (2)	45k (65k)
Claude Opus 5.5	88%	56%	2 (3)	71k (114k)

Table 1: Context and caching over the counted episodes. The narrator's share is lower because each restart re-reads its first frame and the account at full price.

B Tools

Table 2 lists the eighteen tools the actor can call. Every result returns the state of both arms (tool position and angles, gripper gap, what each arm holds), fresh camera images and the narrator's entries since the previous result. On the dual-arm robot, the motion and gripper tools take the name of the arm to move.

group	tool	what it does
perception	`look`	returns fresh images from the head and wrist cameras, with known object ids drawn on them
	`find`	detects every instance of a named object in one camera (SAM 3 on RGB-D) and registers each as an id with a 3D position in the base frame
	`measure`	turns one pixel, or an ordered list of pixels, into base-frame 3D points
motion	`move_above`	moves the gripper to an object or point plus offsets, with given tool angles
	`move_relative`	moves the gripper by (dx, dy, dz), at most 0.15 m per axis
	`descend`	moves straight down to an object's grasp height, to a given height or by a given distance
	`lift`	moves straight up by a given height
	`tilt`	rotates the tool in place: pitch tilts the approach away from vertical, roll spins it about its axis
	`move_until_contact`	moves along a straight line and stops when the joint current indicates contact, read about 25 times per second
	`follow_path`	moves through a list of points in one continuous motion, holding the tool angles
	`home`	returns the arm to its home pose with the gripper open
	`wait`	waits for up to 120 s while the narrator keeps watching, then observes
gripper	`open_gripper`	opens the gripper to a given gap
	`close_gripper`	closes in small increments and stops when the fingers stop; reports whether something is held from where the closure stopped
task	`plan`	records the ordered steps of the task
	`mark_step`	updates one step with the evidence from a new observation
	`done`	ends the episode; refused while a planned step is pending
chain	`do`	runs up to eight of the tools above, sixteen runs in all, with a check and an optional repeat condition on each step (Eq. 1)

Table 2: The actor's tools.

C Simulation-Ready Twins of the Scene

Reins can also rebuild the scene in front of the robot as a MuJoCo simulation from one calibrated head-camera RGB frame. This tool, `model_scene`, was not used in the episodes of Section 3. The model names the objects in the image and gives each a category, a box and any hint it can read off the image, such as an upside-down mug or a hidden handle. The tool returns a scene in the robot's world frame with, for every object, a closed mesh, collision geometry, mass, inertia, joints and surface appearance, together with the calibrated head camera, the bases of both arms and a verdict on whether each object is ready for simulation.

Geometry. Each object is built as a parametric model of its category, by code that drives Blender, rather than fused from depth: a mug is a hollow lathe profile with a swept handle, a drawer cabinet has its drawers on prismatic joints, and a towel or shirt is a deformable sheet; bricks, caps, sponges, brushes and dustpans have builders of their own, and a new category is one more builder. SAM (ViT-H) [37] segments each object from its box, and the mask's footprint on the table gives a first guess of position, axis and size. The logs keep only the head camera's RGB frames, so depth comes from a monocular network, scaled by the table's calibrated height. The models are then fitted to the masks through the calibrated camera by silhouette IoU, with Nelder–Mead over position, yaw and size, and discrete structure, such as a brick's stud count or the side a mug's handle is on, is found by

enumeration. Collision geometry is built for contact: a mug's wall is a ring of 16 thin boxes, so an upturned mug can cover a cap and push it along.

Appearance and physics. Appearance comes from inverse rendering. The scene is rendered in white from the head camera under a uniform sky and under 33 directional lights, and the frame is explained as

$$I(\mathbf{x}) \;\approx\; \boldsymbol{\rho}_{m(\mathbf{x})} \odot \Big(\mathbf{c} \sum_{k=1}^{33} w_k \, S_k(\mathbf{x}) \;+\; \mathbf{s}\, K(\mathbf{x})\Big), \qquad w_k \geq 0, \tag{3}$$

where S_k and K are the renders under light k and under the sky, $\mathbf{c}$ and $\mathbf{s}$ the colours of the lights and of the sky, and $\boldsymbol{\rho}_m$ the albedo of material m. Alternating non-negative least squares solves for all of them on the pixels where model and mask agree, with white balance anchored on a surface known to be white, and the albedo is baked into each object's texture wherever the camera saw it. Mass and inertia come from the closed meshes and material densities, and the scene settles in MuJoCo [38] before it is checked.

Sim-ready gate. Each movable body must pass the checks in Table 3. Collision fidelity is the 95th percentile of the distance from the collision primitives' surface outside the closed mesh to the mesh, with inside and outside decided by the generalized winding number [39]; agreement with the frame is the 90th percentile of the distance between the model's outline and the mask's. A cloth, simulated as a MuJoCo flex, is checked at rest.

check	measure	pass
closed mesh	non-manifold edges; signed volume	0; > 0
collision fidelity	primitives' surface outside the mesh (p95)	≤ 3 mm
stays put	drift during 1–1.5 s of simulation	≤ 5 mm, $\leq 3^\circ$
matches the frame	outline distance to the SAM mask (p90)	≤ 3 px
no initial contact	deepest penetration while settling	≤ 2 mm
cloth, at rest	drift over the last 0.25 s; edge strain (p95); outline distance	≤ 1 mm; $\leq 2\%$; ≤ 3 px

Table 3: The sim-ready gate, applied to each movable body.

Results. We built a twin from the first frame of one episode of each of the five tasks of Section 3 (Table 4, Fig. 3). Of the 14 movable objects, 11 passed the gate: all five in the shell game, three of five in the sweep, one of the two in the drawer scene, and both cloths. The certified objects overlap their masks with IoU 0.90–0.99, their outlines lie 0.7–2.4 px from the models', and their collision geometry stays within 0.2–2.1 mm of their surfaces; the recovered lighting reproduces the frame to within 11–19 of 255 levels. Enumerating stud counts identified the drawer scene's brick as a 6×4 brick and raised its IoU from 0.75 to 0.90, and in the shell game the three mugs, fitted independently, agree in diameter and in height to within 1.3 mm. In every twin a scripted task, driven by mocap hands rather than by the model, runs to completion: a mug covers a cap and pushes it 70 mm, a brush sweeps two caps and a sponge into a held dustpan, a drawer opens 120 mm, takes the brick and closes without moving the cabinet, and the towel and the shirt are folded. Building a scene of rigid objects takes 9–12 min on 12 CPU cores.

task	certified	IoU	outline (px)	light	scripted task in MuJoCo
shell game	5/5	0.92–0.96	0.95–1.64	11.4	cover a cap, push it 70 mm
sweep	3/5	0.95–0.97	0.68–1.37	17.6	sweep 2 caps and a sponge into the pan
drawer	1/2	0.90	1.37	19.0	open 120 mm, place the brick, close
towel	1/1	0.99	1.16	16.2	fold in half (296 $\rightarrow$ 151 mm)
shirt	1/1	0.99	2.39	17.8	fold over (678 $\rightarrow$ 441 mm)

Table 4: Twins of the five tasks of Section 3, each built from one head-camera RGB frame. Certified: objects that pass the gate (a cabinet and its drawers count as one). IoU and outline distance are over the certified objects; light is the mean absolute error of Eq. (3) in 8-bit sRGB levels.

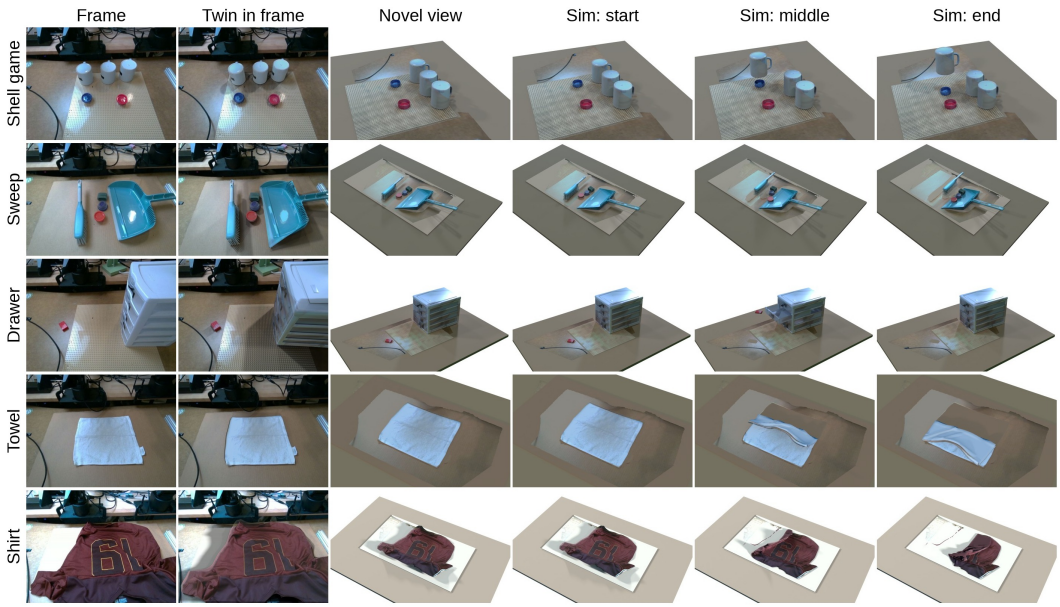


Figure 3: Twins of the five tasks of Section 3, each from one head-camera RGB frame. Columns, from left: the frame; the twin rendered into the frame from the same calibrated camera; the twin from another view; and the start, middle and end of the scripted task in MuJoCo, rendered in Blender.