BARS-RL: BIMANUAL ARTICULATED-OBJECT MANIPULATION UNDER RESISTANCE WITH STABILIZATION

Minjae Kim[*1] **Junhyeok Lee**[*1] **John Prada**[1] **Junhan Jeong**[1] **Chanmi Park**[2] **Yongseok Lee**[†1]
[1]Daegu Gyeongbuk Institute of Science and Technology (DGIST)
[2]Electronics and Telecommunications Research Institute (ETRI)
yslee@dgist.ac.kr

ABSTRACT

Robots in human environments frequently interact with articulated objects that rest unanchored on a surface (e.g., microwaves, toaster ovens), and opening these objects requires overcoming an initial opening resistance. A pulling force below this resistance leaves the moving part closed, whereas one above it can drag the object body, so one arm must pull while the other holds the object body in place. Existing methods that assume anchored objects or negligible opening resistance do not capture this coupling between opening and stabilization. To address these limitations, we propose **BARS-RL**, a reinforcement learning framework in which a single policy allocates force between a task arm and a support arm without observing the resistance or friction. A stabilization reward penalizes contact forces near the base-slip boundary, providing a training signal before the object body moves. Training follows a curriculum organized by the resistance-to-friction ratio, which moves from an anchored object body to an unanchored one and finally mixes episodes that base friction alone can hold with those that require support.

On three simulated objects with revolute and prismatic joints, BARS-RL achieves an average success rate of 83.5%, outperforming baselines, and on real robots it succeeds in 19 of 20 trials across two appliances. Supplementary videos are available at https://frabjous-beijinho-5aaac5.netlify.app/.

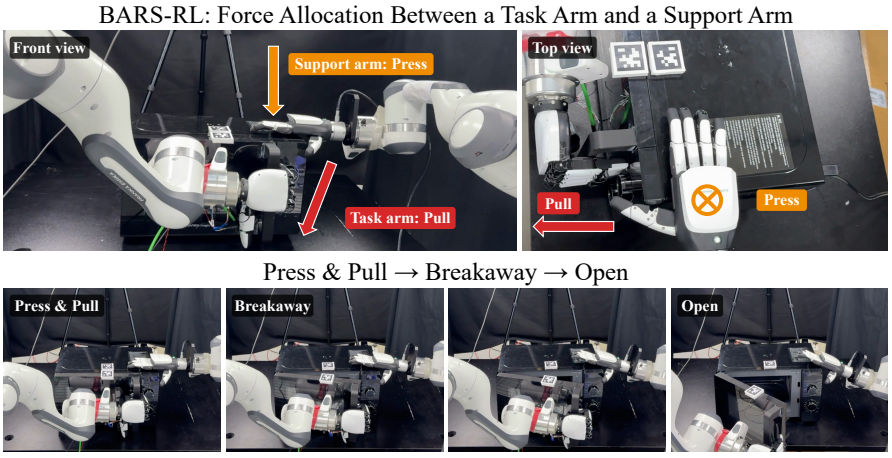


Figure 1: **Conceptual overview of BARS-RL.** BARS-RL allocates force between a task arm that opens the door and a support arm that presses the object to hold it in place, even when the resistance is unknown (top). The two arms press and pull until the closure breaks away, and the task arm then opens the door (bottom).

[*]Equal contribution. [†]Corresponding author.

1 INTRODUCTION

Robots that work in household and industrial settings must manipulate articulated objects such as doors and drawers. Learning-based methods (Mo et al., 2021; Eisner et al., 2022; Wang et al., 2025a) and classical controllers (Jain & Kemp, 2010; Rizzi et al., 2023) have made steady progress on this problem. However, these methods rarely consider articulated objects that rest unanchored on a surface and stay closed until the pull exceeds a breakaway force. In this work, we study how to open such objects (e.g., microwaves, mini fridges, and air fryers) with two arms, one pulling the door open while the other presses the object body to hold it in place.

An important question in this problem is how the two arms should share the force. Existing methods avoid this question through two assumptions. They treat the object body as fixed to the environment or heavy enough not to move, so that the environment absorbs the reaction to the opening force (Rizzi et al., 2023; Mo et al., 2021; Wang et al., 2025a). They also either ignore the opening resistance (Mo et al., 2021; Eisner et al., 2022) or model it as a lock released by a separate action, such as rotating a handle or pressing a button (Li et al., 2024b; Wang et al., 2025b). Without these assumptions, pulling the door can drag the whole object toward the robot before the door opens. A successful policy must therefore satisfy two requirements simultaneously. First, the task arm must pull with a force exceeding the breakaway force. Second, the support arm must press the object body so that base friction can resist this pull.

Both requirements are difficult to satisfy because the robot cannot observe the physical parameters (e.g., resistance and friction) of the object. Our key insight is that a single quantity, the ratio of the opening resistance to the friction, determines whether the support arm is required. Building on this insight, we propose **BARS-RL**, a bimanual reinforcement learning framework in which a single policy allocates force between the task arm and the support arm from force/torque (F/T) and tactile feedback. BARS-RL combines a stabilization reward, which penalizes forces that cause base slip, with a curriculum organized by the resistance-to-friction ratio.

We evaluate BARS-RL in simulation on three articulated objects with revolute and prismatic joints. Without observing the physical parameters, BARS-RL achieves an average success rate of 83.5%, outperforming both scripted and learned baselines and approaching an oracle that observes these parameters, and on real robots it succeeds in 19 of 20 trials across two appliances. To the best of our knowledge, this is the first work to address the opening of unanchored articulated objects under opening resistance, where the robot must overcome the resistance while holding the object body in place.

The contributions of this paper are as follows:

- **Stability quantitative formulation.** We represent latches, detents, and magnetic seals by a single breakaway force and formulate the base-slip margin of an unanchored object body. From this formulation, we derive the resistance-to-friction ratio, which determines whether the support arm is required and is comparable across objects of different mass and friction.
- **Bimanual force allocation.** We propose BARS-RL, in which a single policy allocates force between the task arm and the support arm from proprioceptive, F/T, tactile, and object-state feedback. A stabilization reward penalizes contact forces near the base-slip boundary, and the physical parameters are given only to the critic during training.
- **Physics-organized curriculum and validation.** We organize training by the resistance-to-friction ratio, moving from an anchored object body to an unanchored one before sampling the full range of the ratio. In simulation, BARS-RL outperforms scripted and learned baselines on three articulated objects, and our ablations show that the curriculum is necessary and that organizing it by the ratio improves success. On real robots, it succeeds in 19 of 20 trials across two appliances.

2 RELATED WORK

Articulated object manipulation. Prior work on articulated object manipulation predicts actionable regions from visual input (Mo et al., 2021; Wu et al., 2022; Geng et al., 2023), estimates articulation flow (Eisner et al., 2022; Zhang et al., 2023), reasons about parts with foundation models (Geng et al., 2024; Li et al., 2024a), scales policy learning in simulation (Wang et al., 2025a),

or opens doors and drawers with classical controllers under kinematic uncertainty (Jain & Kemp, 2010; Karayiannidis et al., 2016; Rizzi et al., 2023). These methods are typically developed in simulators whose articulated objects are fixed to the world and whose joints move without resistance once actuated (Xiang et al., 2020). Opening resistance has been addressed by discovering opening strategies through interaction (Wang et al., 2022; Zhang et al., 2025), or by modeling the closure as a latch with friction (Rajeswaran et al., 2018), a lock released by a preceding motion (Li et al., 2024b; Wang et al., 2025b), or a handle unlatched under force feedback (Wang et al., 2025c; Kang et al., 2024). Measurements on everyday objects also show that breakaway forces vary widely (Jain & Kemp, 2013). In all of these works, however, the object body is anchored or treated as stable, so the opening force does not move the object. Tac-Man (Zhao et al., 2025) handles unexpected part motion through tactile feedback, but it does not model when the pull makes an unanchored object body slide. In contrast, we open unanchored objects under an unknown breakaway force, which requires opening the door and holding the object body in place at the same time (Sections 3 and 4.1).

Force-aware and bimanual manipulation. Contact-rich manipulation methods regulate interaction forces through impedance or hybrid force/position control (Do et al., 2025; Fang et al., 2026; Tahmaz et al., 2025) or learn from force-informed demonstrations (Liu et al., 2025b; Ablett et al., 2025). Because the object body is fixed in these works, the applied force only has to move the joint. Bimanual policies assign complementary roles to the two arms (Fu et al., 2025; Wang & Johns, 2025; Lin et al., 2023; Huang et al., 2025), and the closest to ours are those in which one arm stabilizes the object while the other acts on it (Grannen et al., 2023; Liu et al., 2025a; Lin et al., 2025). Cooperative manipulation regulates internal forces between two arms that rigidly grasp a known load (Uchiyama & Dauchez, 1988; Caccavale et al., 2008; Smith et al., 2012). In both lines of work, the stabilized object is grasped or rests on a fixture, so the stabilizing force does not determine whether the task succeeds. In our problem, the support arm presses on an unanchored object body, and this force raises the friction the base can sustain. Using an external contact to raise friction relates our work to extrinsic dexterity (Chavan-Dafle et al., 2014), fixturing under friction limits (Holladay et al., 2024), friction-cone analysis in grasping (Murray et al., 1994), and incipient-slip detection (Veiga et al., 2015; Chen et al., 2018). Hidden physical parameters are often handled through teacher–student distillation or online adaptation (Chen et al., 2020; Lee et al., 2020; Kumar et al., 2021); we instead use an asymmetric actor–critic (Pinto et al., 2018), which gives the resistance and friction only to the critic. To our knowledge, no prior work uses base friction to coordinate two arms that open an articulated object under unknown resistance.

Curriculum learning. Curricula for RL (Bengio et al., 2009; Narvekar et al., 2020; Portelas et al., 2020b) order episodes by start-state or goal difficulty (Florensa et al., 2017; 2018), learning progress (Portelas et al., 2020a), expanding randomization ranges (OpenAI et al., 2019), or regret- and replay-based level selection (Dennis et al., 2020; Jiang et al., 2021). Staged training has also been effective in locomotion (Rudin et al., 2022) and contact-rich manipulation (Huang et al., 2026; Wang et al., 2026). These methods mostly treat difficulty as a generic variable, without tying it to the physics of the task. Our curriculum instead schedules episodes by the resistance-to-friction ratio, which determines whether the support arm is required (Section 4.3).

3 Force Allocation Formulation

We consider opening an unanchored articulated object under opening resistance. The task arm grasps the object handle, the support arm presses on the object body, and the object body rests on the table through a frictional base contact. Let q_t denote the articulation state of the moving part at time t and q_g the target state; the goal is to drive q_t to q_g while the object body stays in place on the table. Each task instance is characterized by a physical parameter vector $\zeta = (F_{\text{res}}, \mu_{\text{b}}, m)$ that the robot cannot observe. Here F_{res} is the breakaway force of the closure that holds the moving part shut, such as a latch or a magnetic seal, measured along the opening direction; μ_{b} is the friction coefficient of the base contact; and m is the total

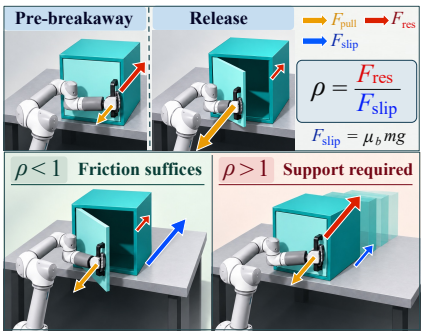


Figure 2: **Resistance-to-friction ratio.** Base friction alone resists the pulling force when $\rho < 1$, whereas the support arm must press on the object body when $\rho > 1$.

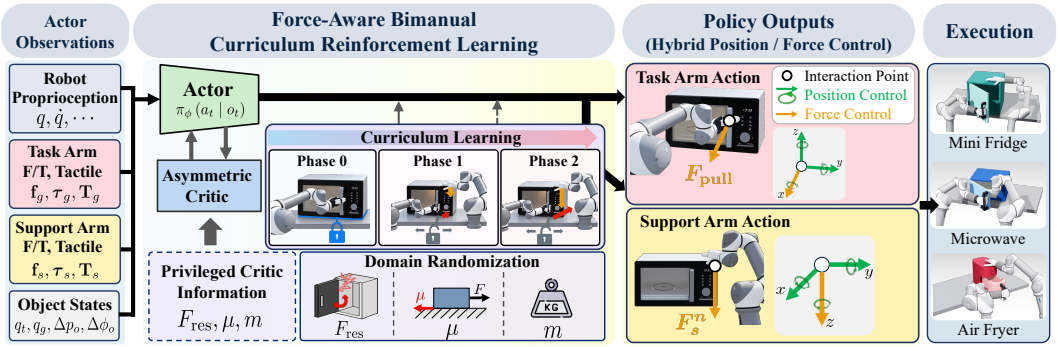


Figure 3: **Pipeline of BARS-RL.** Single policy maps proprioceptive, F/T, tactile, and object-state observations to hybrid position/force commands for both arms. During training, only the critic receives the opening physical parameters, and a curriculum organized by the ρ schedules the episodes.

object mass. We define this task as a *bimanual force allocation* problem: the task arm must apply a force that exceeds a resistance of unknown magnitude, while the support arm presses on the object body to hold it in place.

Base-slip margin and the ratio ρ. While the closure is engaged, the task arm pulls on the handle with force F_{pull} along the opening direction, and the closure transmits this force to the object body. The support arm presses down on the top of the object body with normal force F_{s}^n, which raises the normal load on the base contact to $N_{\text{b}} = mg + F_{\text{s}}^n$. Let F_T denote the component of the net force that the two arms exert on the object along the opening direction; this is the load the base contact must resist. The task arm contributes F_{pull} to F_T, and as long as the support arm does not push the object body along the opening direction, any friction under the support hand opposes it, so $F_T \leq F_{\text{pull}}$. Neglecting tipping and off-axis force, both of which are penalized in the reward (Section 4.2), we define the *base-slip margin* as the difference between the friction capacity of the base contact and this load,

$$\delta_{\text{b}} = \mu_{\text{b}} N_{\text{b}} - F_T. \tag{1}$$

A positive margin means that base friction has capacity to spare, a margin near zero places the object body at the slip boundary, and a negative margin means that the object body slides.

Breakaway occurs once F_{pull} exceeds F_{res} (Section 4.1). Since $F_T \leq F_{\text{pull}}$, the margin stays positive until breakaway if

$$F_{\text{res}} < \mu_{\text{b}} \left(mg + F_{\text{s}}^n \right), \tag{2}$$

which is a sufficient force-allocation condition for the two arms: the task arm must reach F_{res}, and the support arm must press so that base friction can resist this force. Without support ($F_{\text{s}}^n = 0$), the right-hand side reduces to the slip threshold $F_{\text{slip}} = \mu_{\text{b}} mg$, which defines the resistance-to-friction ratio $\rho = F_{\text{res}}/F_{\text{slip}}$. When $\rho < 1$, base friction alone can hold the object body, and single-arm opening is feasible in principle. When $\rho > 1$, F_{pull} exceeds F_{slip} before breakaway, and equation 2 holds if the support arm presses with

$$F_{\text{s}}^n > mg\,(\rho - 1). \tag{3}$$

This bound depends on the unobserved parameters ζ and is conservative, since it ignores the friction under the support hand that lowers F_T. Because ρ summarizes how much support a task instance requires in a single quantity comparable across objects of different mass and friction, we use $\rho = 1$ as the boundary between the two regimes and organize the curriculum around it (Section 4.3).

4 Method

Overview. We propose BARS-RL (Figure 3), a reinforcement learning framework for opening unanchored articulated objects under opening resistance. Section 4.1 introduces the opening-resistance model, which reduces every closure to a single breakaway force, and Section 4.2 adds stabilization terms built around the base-slip margin to the task reward. For $\rho < 1$, this reward is sufficient on its own. For $\rho > 1$, however, the policy must coordinate both arms, and this coordination is difficult to learn directly; we therefore order training by ρ using the curriculum of Section 4.3.

4.1 Opening Resistance Model

Prior simulation environments model a closure as a kinematic lock that is released by turning a handle or moving a preceding joint, after which the moving part moves freely (Li et al., 2024b; Wang et al., 2025b). These models omit the force that must be overcome before the moving part can move, and it is this force that makes the support arm necessary. We therefore represent latches, detents, and magnetic seals by a single breakaway force F_{res}. While the closure is engaged, a fixed joint attaches the moving part to the object body, so the full pulling force is transmitted to the object body, as assumed in equation 1. At each control step, the wrist force of the task arm is projected onto the opening direction,

$$F_{\text{pull}}(t) = \left\langle \bar{\mathbf{f}}_{\text{w}}(t),\ \hat{\mathbf{a}}(q_t) \right\rangle, \tag{4}$$

where $\bar{\mathbf{f}}_{\text{w}}(t)$ is the wrist F/T force of the task arm averaged over a short moving window and $\hat{\mathbf{a}}(q_t)$ is the unit opening direction. Once $F_{\text{pull}}(t)$ exceeds F_{res}, the fixed joint is released, and the moving part moves under the stiffness and damping of its own joint. Breakaway is thus modeled as a discrete event, consistent with the abrupt release of most household closures. In addition, F_{res} can be set independently in every episode, which the curriculum in Section 4.3 relies on.

4.2 Manipulation under Resistance with Stabilization

Task reward. The task reward drives the moving part to the target and stops it there,

$$r_{\text{task}} = \lambda_{\text{prog}}\, r_{\text{prog}} + \lambda_{\text{succ}}\, r_{\text{succ}} + \lambda_{\text{pull}}\, r_{\text{pull}} + \lambda_{\text{settle}}\, r_{\text{settle}} - \lambda_{\text{time}} - \lambda_{\text{act}}\, c_{\text{act}}, \tag{5}$$

where r_{prog} is the normalized reduction in distance to the target, credited only while the task arm holds the handle, and r_{succ} is a bonus given when the moving part rests at the target with the grasp held and the object body in place. The shaping terms r_{pull} and r_{settle} reward pulling along the opening direction before the closure releases and coming to rest at the target, respectively; the last two terms penalize elapsed time and action rate. However, none of the dense terms can distinguish opening the moving part from dragging the object body toward the robot.

Base-slip cost. The central stabilization term acts on the base-slip margin δ_{b} of equation 1. We express this margin as the fraction of the base friction capacity demanded by the load, $u_{\text{b}} = F_T/(\mu_{\text{b}} N_{\text{b}})$, and penalize it beyond a safety factor $\kappa < 1$,

$$c_{\text{base}} = \text{clip}\left(\frac{u_{\text{b}} - \kappa}{1 - \kappa},\ 0,\ 1 \right). \tag{6}$$

Without support, u_{b} reaches ρ at breakaway; the cost is therefore active whenever $\rho > \kappa$ unless the support arm presses on the object body, which raises N_{b} as in equation 2. We also include a displacement cost c_{body} for translating and rotating the object body. While this cost responds only after the object body moves, c_{base} responds to the force state of the contact before any displacement accumulates. Because c_{base} requires μ_{b} and the base-contact force, it is used only during training, and neither quantity is given to the actor.

Support reward. The support reward couples the two arms explicitly,

$$r_{\text{sup}} = \mathbf{1}_{\text{pulling}}\, \mathbf{1}_{\text{pressing}}\, \exp\left[-\left(\|\Delta p_o\| / \sigma_o \right)^2 \right], \tag{7}$$

where Δp_o is the translation of the object body since reset, $\mathbf{1}_{\text{pulling}}$ is active while the task arm holds the handle and pulls along the opening direction, and $\mathbf{1}_{\text{pressing}}$ is active while the support arm presses above a threshold. Because both indicators must be active, an idle policy receives no support reward for an object that is merely stationary.

Overall reward. The per-step reward is

$$\begin{aligned} r_t = r_{\text{task}} &+ w_{\text{sup}}\, \lambda_{\text{sup}}\, r_{\text{sup}} - \lambda_{\text{body}}\, c_{\text{body}} \\ &- w_{\text{stab}} \left(\lambda_{\text{base}}\, c_{\text{base}} + \lambda_{\text{overgrip}}\, c_{\text{overgrip}} + \lambda_{\text{offaxis}}\, c_{\text{offaxis}} \right), \end{aligned} \tag{8}$$

where c_{overgrip} penalizes excess grip force, c_{offaxis} penalizes wrist force that does not contribute to opening, and $w_{\text{stab}} \in [0, 1]$ and $w_{\text{sup}} \in \{0, 1\}$ are stage-dependent weights (Section 4.3). The off-axis force and the tipping that the base-slip margin neglects are covered by c_{offaxis} and by the rotation part of c_{body}, respectively. All term definitions, weights, and thresholds are given in Appendix D.3.

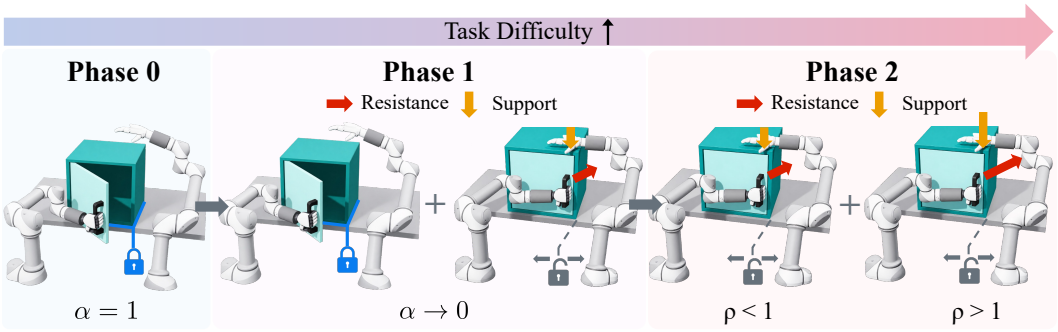


Figure 4: **Curriculum phases.** Training proceeds from anchored articulation (Phase 0) to a mixture of anchored and unanchored episodes with $\rho < 1$ (Phase 1), and then to a single stage that samples ρ uniformly over an interval straddling $\rho = 1$, in which episodes with $\rho > 1$ dominate (Phase 2).

4.3 Physics-Organized ρ-Curriculum

Motivation. Because most episodes in the final task distribution have $\rho > 1$, optimizing equation 8 directly on this distribution requires the policy to discover pulling and pressing at the same time; until it does, pulling only drags the object body and the closure never releases. The ratio ρ separates episodes by this difficulty at $\rho = 1$: below one, base friction absorbs the resistance and a single arm suffices; above one, the object body starts to slide before the pulling force reaches the breakaway force, so the closure does not release unless the support arm presses on the object body. It is not, however, a property of the resistance alone: the same F_{res} is a single-arm task on a heavy object with a rough base and a two-arm task on a light, slippery one. We therefore organize the curriculum by ρ itself. The mass m and base friction μ_b are sampled per environment, ρ is sampled per episode from a stage-dependent interval $\mathcal{R}_s$, and the resistance is set to the value that realizes the drawn ratio,

$$F_{\text{res}} = \rho\, F_{\text{slip}}(m, \mu_b). \tag{9}$$

Curriculum phases. Each stage s fixes the probability α_s that the object body is anchored, together with the interval $\mathcal{R}_s$; the stages are grouped into the three phases of Figure 4.

Phase 0: anchored articulation. The object body is rigidly anchored ($\alpha_0 = 1$), the closure offers no resistance, and the contact costs and the support reward are switched off ($w_{\text{stab}} = w_{\text{sup}} = 0$). The policy learns to move the moving part and stop at q_g before it is asked to stabilize anything.

Phase 1: from anchored to unanchored. Each episode is anchored with probability α_s and unanchored otherwise, and α_s is lowered stage by stage to zero. The support reward is enabled ($w_{\text{sup}} = 1$) and the contact costs enter at half weight ($w_{\text{stab}} = 0.5$), while $\mathcal{R}_s$ is held fixed and lies entirely below one. The policy therefore learns to pull without moving an unanchored object body before it encounters a resistance that base friction cannot absorb.

Algorithm 1 ρ-curriculum

Require: Stages $\{(\alpha_s, \mathcal{R}_s, \tau_s)\}_{s=0}^{S}$
$s \leftarrow 0$, $\mathcal{B} \leftarrow \emptyset$
for each PPO iteration **do**
  for each environment where episode is done **do**
    if episode began under stage s **then**
      Append success indicator to $\mathcal{B}$
    end if
    Anchor the object body with probability α_s
    Sample $\rho \sim \mathcal{U}(\mathcal{R}_s)$
    Set $F_{\text{res}} \leftarrow \rho \cdot F_{\text{slip}}(m, \mu_b)$
  end for
  if $|\mathcal{B}| \geq$ number of environments **then**
    if $\text{mean}(\mathcal{B}) \geq \tau_s$ **and** $s < S$ **then**
      $s \leftarrow s + 1$
      Resample m, μ_b
    end if
    $\mathcal{B} \leftarrow \emptyset$
  end if
end for

Phase 2: full ρ range. All episodes are unanchored and the contact costs are at full weight ($w_{\text{stab}} = 1$). A single final stage samples ρ uniformly from an interval $\mathcal{R}_S$ that spans both $\rho < 1$ and $\rho > 1$; the upper end of the μ_b range is clipped so that the largest resistance stays within the wrench limit of the task arm. Phase 2 thus retains the episodes that base friction can hold and adds those that require support, so the policy learns from one distribution both when support is needed and how much force to apply.

Table 1: **Main results.** Success rate (%) on three unanchored articulated objects, reported separately for $\rho \in [0, 1.25)$ and $\rho \in [1.25, 2.0]$.

Method	Microwave		Mini fridge		Air fryer		Avg.	
	$\rho < 1.25$	$\rho \geq 1.25$	$\rho < 1.25$	$\rho \geq 1.25$	$\rho < 1.25$	$\rho \geq 1.25$	$\rho < 1.25$	$\rho \geq 1.25$
ACT	47.3	33.5	27.7	9.8	20.6	11.7	31.9	18.3
ACT w/ F/T	50.9	30.0	30.8	8.6	24.1	11.9	35.3	16.8
ACT w/ F/T w/ Tactile	48.0	18.4	25.4	8.2	24.6	9.4	32.7	12.0
Diffusion Policy	50.9	28.7	37.6	15.4	37.0	21.6	41.8	21.9
Diffusion Policy w/ F/T	66.6	38.2	51.1	14.8	75.0	60.1	64.2	37.7
Diffusion Policy w/ F/T w/ Tactile	58.6	22.7	45.7	19.9	79.3	53.9	61.2	32.2
VoxAct-B	69.9	40.6	26.2	12.1	18.4	14.5	38.2	22.4
VoxAct-B w/ F/T	75.0	62.1	44.1	21.9	55.9	51.2	58.3	45.1
VoxAct-B w/ F/T w/ Tactile	70.7	54.7	27.0	11.7	37.9	33.6	45.2	33.3
Fixed-force bimanual controller	32.0$\pm$1.8	15.7$\pm$0.8	61.5$\pm$2.9	38.1$\pm$2.4	26.1$\pm$1.2	17.7$\pm$1.7	39.9	23.8
Single-arm	**98.8**$\pm$0.4	39.5$\pm$2.5	**98.8**$\pm$0.5	31.8$\pm$1.2	93.5$\pm$1.8	17.1$\pm$1.3	**97.0**	29.5
BARS-RL	91.9$\pm$1.1	**85.3**$\pm$2.0	81.2$\pm$4.1	**68.4**$\pm$3.3	**97.3**$\pm$0.7	**96.7**$\pm$0.6	90.1	**83.5**
Oracle-physics actor (ζ observed)	94.6$\pm$2.3	94.7$\pm$2.8	86.2$\pm$2.5	77.2$\pm$5.5	98.3$\pm$1.1	98.2$\pm$1.4	93.0	90.0

5 Experiments

In this section, we evaluate whether BARS-RL can open unanchored articulated objects under unknown opening resistance while holding the object body in place. Section 5.1 describes the objects, robot, and baselines and compares BARS-RL with them in simulation, Section 5.2 validates it on real robots, and Section 5.3 examines the breakaway model, the support arm, the curriculum, the base-slip cost, and new objects. Unless stated otherwise, we trained each RL method with 5 random seeds and evaluated the checkpoint with the highest training success rate of each seed on 1,024 parallel environments, with ρ sampled uniformly from $[1.25, 2.0]$, a range in which base friction alone cannot hold the object body; Table 1 additionally reports $\rho \in [0, 1.25)$. An episode succeeds when Eq. 18 holds, and we report success rates (%) with standard deviations across seeds.

5.1 Experimental Results

Environment. We evaluated on three unanchored articulated objects in Isaac Lab (NVIDIA et al., 2025): a microwave and a mini fridge with revolute doors, and an air fryer with a prismatic basket. Each object rested on a table without being fixed, and its closure was equipped with the opening-resistance model of Section 4.1. The robot consisted of two RB5-850E (Rainbow Robotics, 2026) arms; the task arm carries an XHand1 (ROBOTERA, 2025) that grasps the handle, and the support arm carries an Inspire hand (Beijing Inspire-Robots Technology Co., Ltd., 2025) that presses on the object body. Observation and action spaces, reward weights, hyperparameters, randomization ranges, and episode reset conditions are given in Appendix D.

Methods. We trained one BARS-RL policy per object with PPO (Schulman et al., 2017; Schwarke et al., 2025) and an asymmetric actor–critic (Pinto et al., 2018). The actor observed only proprioception, wrist F/T, tactile signals, and the object state, whereas the physical parameters ζ were randomized during training and provided to the critic alone.

For imitation learning baselines, we compared ACT (Zhao et al., 2023), Diffusion Policy (Chi et al., 2025), and VoxAct-B (Liu et al., 2025a), each with vision only, with wrist F/T, and with wrist F/T and tactile signals. For each object, we collected 100 demonstrations with ρ sampled from $[0, 2]$, using a scripted expert that observes ζ and the closure state. The policies observed images from a front camera and two wrist cameras, the 24 arm and hand joint positions, and the goal q_g, plus the force and tactile inputs of each variant (details in Appendix D.7).

We also compared against the *fixed-force bimanual controller*, a scripted controller that pulled and pressed with constant force targets, and a single-arm baseline. The fixed-force controller pulled with the largest breakaway force in the final-phase range and pressed with the force that Eq. 3 requires for this breakaway force, plus a 25% margin, within the force limit of the support arm. *Single-arm* was trained with the same objective and curriculum as BARS-RL but without the support arm. As an upper bound, we also report an *oracle-physics actor*, identical to BARS-RL except that its actor also observes ζ.

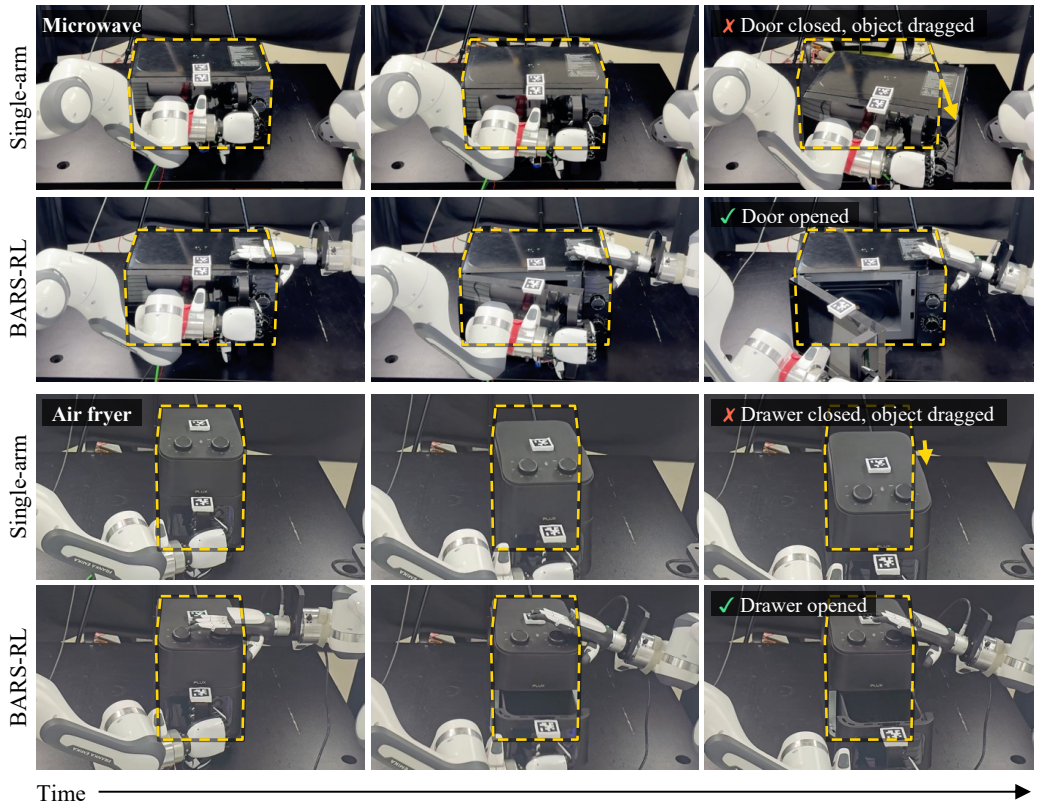


Figure 5: **Real-world rollouts on the microwave and the air fryer.** Each row shows one method from left to right over time, and the dashed outline marks the initial position of the object body. The single-arm policy drags the object body while the door or basket stays closed, whereas BARS-RL presses on the object body with the support arm and opens it.

Results. Table 1 reports the success rate of each method on the three objects. For $\rho \geq 1.25$, a range in which base friction alone cannot hold the object body, BARS-RL achieves an average success rate of 83.5%, outperforming all baselines by more than 38 points on average and reaching about 93% of the oracle-physics actor (90.0%), which directly observes ζ. On the air fryer, BARS-RL (96.7%) nearly matches the oracle (98.2%), whereas the gap is largest on the mini fridge (68.4% vs. 77.2%). Since the oracle differs from BARS-RL only in observing ζ, this gap quantifies the benefit of direct access to the physical parameters over acting from feedback alone. For $\rho < 1.25$, the single-arm policy achieves the highest average success rate (97.0%), since a single arm suffices in this range (Section 5.3); BARS-RL reaches 90.1%, close to the oracle (93.0%).

The imitation learning baselines fall well below BARS-RL in both ranges. Among the vision-only baselines (ACT, Diffusion Policy, and VoxAct-B), the highest average success rates are 41.8% for $\rho < 1.25$ and 22.4% for $\rho \geq 1.25$. While the door remains closed and stationary, visual observations change little with pulling force, making it difficult to judge how much more force to apply. F/T input improves Diffusion Policy and VoxAct-B, raising the best baseline rates to 64.2% and 45.1% in the two ranges, respectively, with both rates still below those of BARS-RL. Adding tactile signals does not improve the average rates over F/T alone: for $\rho \geq 1.25$, ACT reaches 12.0%, Diffusion Policy 32.2%, and VoxAct-B 33.3%. The 100 training demonstrations may cover too few combinations of friction and latch resistance for the policies to coordinate pulling and stabilizing forces in unseen conditions, even with F/T feedback. We further analyze failures of the single-arm policy and BARS-RL without curriculum in Section 5.3 and Appendix C.

5.2 Real-World Experiments

Setup. To test whether a policy trained in simulation works on real robots, we ran real-world experiments on the microwave and the air fryer. Instead of the RB5-850E arms used in simulation,

we used two Franka arms with joint torque sensors, because an external six-axis F/T sensor mounted on the wrist could not measure the pulling force accurately due to crosstalk between its axes (Shin et al., 2025; Yu et al., 2021). The task arm is a Franka FR3 with an XHand1, and the support arm is a Franka Panda with an Inspire hand. We estimated the end-effector wrench from the joint torques with a momentum observer (Haddadin et al., 2017). We attached markers to the object body and the handle and tracked them with one top-view Intel RealSense D456 camera to obtain the pose of the object body and the articulation state q_t. The policy was trained in simulation as in Section 4, with only the robot model replaced by the Franka arms, and then deployed on the real robots. The initial poses of both arms were set by the procedure of Appendix D.1.

Results. We ran 10 trials per method on each object and determined success by the same condition as in simulation (Eq. 18). BARS-RL succeeded in 9 of 10 trials on the microwave and in all 10 trials on the air fryer, whereas the single-arm policy did not succeed in any trial (Table 2). Figure 5 shows rollouts on both objects. With the single-arm policy, the object body is dragged before the closure releases, so the door (basket) stays closed. BARS-RL, however, opens the door (basket) while the object stays stable, which matches the behavior observed in simulation (Section 5.3).

Table 2: **Real-world success.**

Method	Microwave	Air fryer
Single-arm	0/10	0/10
BARS-RL	**9/10**	**10/10**

5.3 Q&As

We examine the formulation and the curriculum design of BARS-RL through the following Q&As.

Q: How closely does the breakaway model match real closures?

A: The breakaway model reproduces the behavior of real closures, where the force builds up gradually and is released abruptly. To test this, we slowly pulled the handles of the three real appliances with a robot arm while estimating the pull force from the joint torques (Haddadin et al., 2017), and replayed the same pull in simulation with the breakaway model, whose breakaway force was set to the measured value, repeating each pull five times in both settings (Figure 6). For all three objects, the measured pull force increases until it reaches the breakaway force and then drops sharply, and the simulated force shows the same pattern. The same holds for the magnetically latched mini fridge. Although its magnetic holding force decays continuously with the gap rather than vanishing at once, the door snaps open at release, so the force drops as abruptly as with the mechanical latches of the microwave and air fryer.

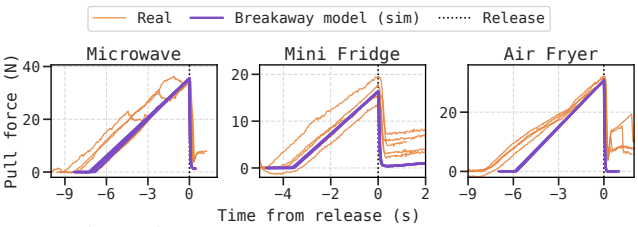


Figure 6: **Breakaway model vs. real closures**

Q: Is the support arm necessary?

A: In our formulation (Section 3), ρ is the ratio of the opening resistance to the friction the base can sustain on its own, so the object body begins to slide once $\rho > 1$ and the support arm is needed to hold it in place. To test this, we trained a single-arm policy for each object under the same settings as BARS-RL and evaluated both policies over $\rho \in [0.75, 2.0]$ (Figure 7). At low ρ, the single-arm policy performs better (e.g., 98% vs. 76% on the mini fridge at $\rho = 0.75$). Since a single arm suffices in this range and the success rate of BARS-RL changes little with ρ, we believe this gap is a side effect of using both arms even when support is unnecessary. As ρ increases, the two policies become comparable at $\rho = 1.25$, where the object body already slides but its displacement mostly stays within the success tolerance. The gap becomes clear at $\rho = 1.5$, where the displacement exceeds the tolerance, and the single-arm policy drops to 11% on the microwave and 0% on the other two objects, while BARS-

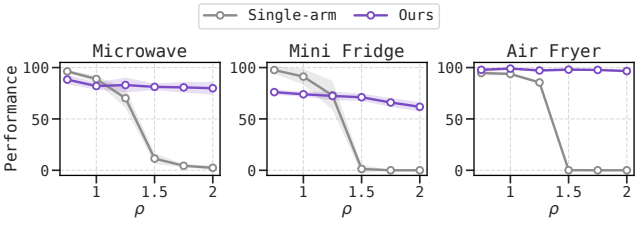


Figure 7: **Support arm becomes necessary at high ρ.**

RL maintains its success rate over the whole range. All single-arm failures in this regime are drag failures, which correspond to the event $\delta_b < 0$ that ρ predicts, and the same pattern holds across objects with different mass, friction, and joint type. The support arm therefore becomes necessary once the object body slides beyond the task tolerance, which occurs between $\rho = 1.25$ and $\rho = 1.5$ for all three objects.

Q: Is a curriculum necessary, and why define it in terms of ρ?

A: BARS-RL trains on anchored objects, then on unanchored objects with $\rho < 1$, and finally on a range of ρ that spans both $\rho < 1$ and $\rho > 1$ (Section 4.3). To test this design, we compared it with two variants: *w/o curriculum*, which was trained directly on the final-phase distribution, and an F_{res} *curriculum*, which shared Phases 0 and 1 but raised the opening resistance F_{res} stage by stage in the final phase instead of scheduling by ρ. Table 3 shows the results. Without the curriculum, the policy fails on every object, because it must learn to pull and press at the same time from the start. The F_{res} curriculum learns the task but falls behind BARS-RL on all three objects (68.6% vs. 83.5% on average). We believe this is because the same F_{res} corresponds to different values of ρ under the randomized mass and friction, so each stage mixes episodes that need no support with episodes that need strong support. In contrast, ρ directly measures how much support an episode requires, as shown by the consistent transition across objects in Figure 7.

Table 3: **Effect of curriculum learning design.**

Training	Microwave	Mini fridge	Air fryer	Avg.
w/o curriculum	0.0	0.0	0.0	0.0
F_{res} curriculum	53.9±4.8	64.2±2.5	87.7±6.6	68.6
BARS-RL	**85.3**±2.0	**68.4**±3.3	**96.7**±0.6	**83.5**

Q: Does the base-slip cost matter?

A: Yes, removing c_{base} substantially degrades the policy, lowering the average success rate from 83.5% to 59.9% (Table 4). This variant removed only c_{base} and kept c_{body} and all other terms unchanged. The success rate drops on all three objects, most on the mini fridge (68.4% to 31.7%) and the microwave (85.3% to 61.8%), and by 10.5 points even on the air fryer (96.7% to 86.2%). The standard deviation across seeds also increases on every object, indicating that training without c_{base} is less stable. While c_{body} responds only after the object body moves, c_{base} directly penalizes force states approaching the base-slip boundary before displacement accumulates. Removing c_{base} therefore removes this explicit pre-slip penalty, while the remaining reward terms are retained.

Table 4: **Effect of the base-slip cost.**

Method	Microwave	Mini fridge	Air fryer	Avg.
w/o c_{base}	61.8±8.0	31.7±7.6	86.2±2.1	59.9
BARS-RL	**85.3**±2.0	**68.4**±3.3	**96.7**±0.6	**83.5**

Q: Does BARS-RL apply to new objects without object-specific design?

A: Yes, BARS-RL achieves success rates above 92% in both ranges of ρ on two new objects, with the reward, curriculum, and hyperparameters unchanged (Table 5). To test this, we trained BARS-RL on a cabinet with a prismatic drawer and an oven with a revolute door from ArtVIP (Jin et al., 2026). The initial poses of both arms were computed from the object geometry by the procedure of Appendix D.1, and all other settings were identical to those of the three objects above. Given a new object, BARS-RL can therefore be applied by training alone, without redesigning the reward or hand-tuning the initial poses.

Table 5: **Evaluation on additional objects.** Success rate (%) over three seeds.

Object	$\rho < 1.25$	$\rho \geq 1.25$
Cabinet (prismatic)	94.2±1.1	92.7±1.6
Oven (revolute)	96.6±0.4	92.2±1.2

6 Conclusion

We present BARS-RL, a reinforcement learning framework in which a bimanual policy allocates force between a task arm and a support arm to open unanchored articulated objects under unknown opening resistance. Through the base-slip margin, we show that the resistance-to-friction ratio determines when the support arm is required, and we use this ratio to organize the training curriculum. Our experiments on three simulated objects demonstrate that BARS-RL outperforms scripted and learned baselines and approaches an oracle that observes the physical parameters, and on real robots it succeeds in 19 of 20 trials across two appliances.

REFERENCES

T. Ablett, O. Limoyo, A. Sigal, A. Jilani, J. Kelly, K. Siddiqi, F. Hogan, and G. Dudek. Multimodal and force-matched imitation learning with a see-through visuotactile sensor. *IEEE Transactions on Robotics*, 41:946–959, 2025.

Beijing Inspire-Robots Technology Co., Ltd. Inspire hand official website. https://inspire-robots.store/collections/the-dexterous-hands, 2025. Accessed: 2025-03-17.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009.

Fabrizio Caccavale, Pasquale Chiacchio, Alessandro Marino, and Luigi Villani. Six-DOF impedance control of dual-arm cooperative manipulators. *IEEE/ASME Transactions on Mechatronics*, 13(5): 576–586, 2008.

Nikhil Chavan-Dafle, Alberto Rodriguez, Robert Paolini, Bowei Tang, Siddhartha S. Srinivasa, Michael Erdmann, Matthew T. Mason, Ivan Lundberg, Harald Staab, and Thomas Fuhlbrigge. Extrinsic dexterity: In-hand manipulation with external forces. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1578–1585, 2014.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*, pp. 66–75, 2020.

Wei Chen, Heba Khamis, Ingvars Birznieks, Nathan F. Lepora, and Stephen J. Redmond. Tactile sensors for friction estimation and incipient slip detection—toward dexterous robotic manipulation: A review. *IEEE Sensors Journal*, 18(22):9049–9064, 2018.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations (ICLR)*, 2016.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

T.-D. Do, N. Gireesh, J. Wang, and H. Wang. Watch less, feel more: Sim-to-real RL for generalizable articulated object manipulation via motion adaptation and impedance control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4806–4812, 2025.

B. Eisner, H. Zhang, and D. Held. FlowBot3D: Learning 3D articulation flow to manipulate articulated objects. In *Robotics: Science and Systems (RSS)*, 2022.

H. Fang, S. Tang, M. Mei, H. Qin, Z. He, J. Chen, Y. Feng, C. Wang, W. Liu, Z. He, C. Lu, and S. Wang. Force Policy: Learning hybrid force-position control policy under interaction frame for contact-rich manipulation. In *Robotics: Science and Systems (RSS)*, 2026.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2017.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International Conference on Machine Learning (ICML)*, 2018.

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4066–4083, 2025.

H. Geng, Z. Li, Y. Geng, J. Chen, H. Dong, and H. Wang. PartManip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2978–2988, 2023.

H. Geng, S. Wei, C. Deng, B. Shen, H. Wang, and L. J. Guibas. SAGE: Bridging semantic and actionable parts for generalizable articulated-object manipulation under language instructions. In *Robotics: Science and Systems (RSS)*, 2024.

Suresh Goyal, Andy Ruina, and Jim Papadopoulos. Planar sliding with dry friction. Part 1. Limit surface and moment function. *Wear*, 143(2):307–330, 1991. doi: 10.1016/0043-1648(91)90104-3.

Jennifer Grannen, Yilin Wu, Brandon Vu, and Dorsa Sadigh. Stabilize to act: Learning to coordinate for bimanual manipulation. In *Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pp. 563–576, 2023.

Sami Haddadin, Alessandro De Luca, and Alin Albu-Schäffer. Robot collisions: A survey on detection, isolation, and identification. *IEEE Transactions on Robotics*, 33(6):1292–1312, 2017.

Rachel Holladay, Tomás Lozano-Pérez, and Alberto Rodriguez. Robust planning for multi-stage forceful manipulation. *The International Journal of Robotics Research*, 43(3):330–353, 2024.

Robert D. Howe and Mark R. Cutkosky. Practical force-motion models for sliding manipulation. *The International Journal of Robotics Research*, 15(6):557–572, 1996. doi: 10.1177/027836499601500603.

B. Huang, J. Xu, I. Akinola, W. Yang, B. Sundaralingam, R. O'Flaherty, D. Fox, X. Wang, A. Mousavian, Y.-W. Chao, and Y. Li. VT-Refine: Learning bimanual assembly with visuotactile feedback via simulation fine-tuning. In *Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2484–2500, 2025.

C.-Y. Huang, A. Efendi, and Y.-C. Wang. Curriculum-guided deep reinforcement learning with fuzzy rewards for autonomous push-grasp manipulation in cluttered environments. *Robotics and Autonomous Systems*, 198:105325, 2026.

Advait Jain and Charles C. Kemp. Pulling open doors and drawers: Coordinating an omnidirectional base and a compliant arm with equilibrium point control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

Advait Jain and Charles C. Kemp. Improving robot manipulation with data-driven object-centric models of everyday forces. *Autonomous Robots*, 35(2–3):143–159, 2013.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *International Conference on Machine Learning (ICML)*, 2021.

Zhao Jin, Zhengping Che, Tao Li, Zhen Zhao, Kun Wu, Yuheng Zhang, Yinuo Zhao, Zehui Liu, Qiang Zhang, Xiaozhu Ju, et al. Artvip: Articulated digital assets of visual realism, modular interaction, and physical fidelity for robot learning. In *International Conference on Learning Representations*, volume 2026, pp. 75272–75301, 2026.

G. Kang, H. Seong, D. Lee, and D. H. Shim. A versatile door opening system with mobile manipulator through adaptive position-force control and reinforcement learning. *Robotics and Autonomous Systems*, 180:104760, 2024.

Yiannis Karayiannidis, Christian Smith, Francisco Eli Viña Barrientos, Petter Ögren, and Danica Kragic. An adaptive control approach for opening doors and drawers under uncertainties. *IEEE Transactions on Robotics*, 32(1):161–175, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Charles A. Klein and Bruce E. Blaho. Dexterity measures for the design and control of kinematically redundant manipulators. *The International Journal of Robotics Research*, 6(2):72–83, 1987.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems (RSS)*, 2021.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020.

X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong. ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18061–18070, 2024a.

Yu Li, Xiaojie Zhang, Ruihai Wu, Zilong Zhang, Yiran Geng, Hao Dong, and Zhaofeng He. UniDoorManip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments. *arXiv preprint arXiv:2403.02604*, 2024b.

Toru Lin, Zhao-Heng Yin, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Twisting lids off with two hands. In *Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 5220–5235, 2025.

Y. Lin, A. Church, M. Yang, H. Li, J. Lloyd, D. Zhang, and N. F. Lepora. Bi-Touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning. *IEEE Robotics and Automation Letters*, 8(9):5472–5479, 2023.

I.-C. A. Liu, S. He, D. Seita, and G. S. Sukhatme. VoxAct-B: Voxel-based acting and stabilizing policy for bimanual manipulation. In *Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4354–4370, 2025a.

W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu. ForceMimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1105–1112, 2025b.

K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6813–6823, 2021.

R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Boca Raton, FL, 1994.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020.

NVIDIA, Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving Rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Robotics: Science and Systems (RSS)*, 2018.

Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments. In *Conference on Robot Learning (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*, pp. 835–853, 2020a.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep RL: A short survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020b.

Rainbow Robotics. RB5-850 collaborative robot official website. https://rainbow-robotics.com/en/products/rb5-850/, 2026. Accessed: 2026-09-21.

A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Robotics: Science and Systems (RSS)*, 2018.

Giuseppe Rizzi, Jen Jen Chung, Abel Gawel, Lionel Ott, Marco Tognon, and Roland Siegwart. Robust sampling-based control of mobile manipulators for interaction with articulated objects. *IEEE Transactions on Robotics*, 39(3):1929–1946, 2023.

ROBOTERA. XHAND1 official website. https://www.robotera.com/en/goods1/4.html, 2025. Accessed: 2025-03-17.

Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning (CoRL)*, volume 164 of *Proceedings of Machine Learning Research*, pp. 91–100, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Clemens Schwarke, Mayank Mittal, Nikita Rudin, David Hoeller, and Marco Hutter. RSL-RL: A learning library for robotics research. *arXiv preprint arXiv:2509.10771*, 2025.

Jiyou Shin, Jinjae Shin, Hong-ryul Jung, Jaeseok Won, Eugene Auh, and Hyungpil Moon. On-site implementation of external wrench measurement via non-linear optimization in six-axis force–torque sensor calibration and crosstalk compensation. *Applied Sciences*, 15(3):1510, 2025.

Christian Smith, Yiannis Karayiannidis, Lazaros Nalpantidis, Xavi Gratal, Peng Qi, Dimos V. Dimarogonas, and Danica Kragic. Dual arm manipulation—a survey. *Robotics and Autonomous Systems*, 60(10):1340–1353, 2012.

A. B. Tahmaz, R. Prakash, and J. Kober. Impedance primitive-augmented hierarchical reinforcement learning for sequential tasks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10973–10979, 2025.

Masaru Uchiyama and Pierre Dauchez. A symmetric hybrid position/force control scheme for the coordination of two robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 350–356, 1988.

Filipe Veiga, Herke van Hoof, Jan Peters, and Tucker Hermans. Stabilizing novel objects by learning to predict tactile slip. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5065–5072, 2015.

Y. Wang and E. Johns. One-shot dual-arm imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5660–5668, 2025.

Y. Wang, R. Wu, K. Mo, J. Ke, Q. Fan, L. J. Guibas, and H. Dong. AdaAfford: Learning to adapt manipulation affordance for 3D articulated objects via few-shot interactions. In *European Conference on Computer Vision (ECCV)*, pp. 90–107, 2022.

Y. Wang, Z. Wang, M. Nakura, P. Bhowal, C.-L. Kuo, Y.-T. Chen, Z. Erickson, and D. Held. ArticuBot: Learning universal articulated object manipulation policy via large scale simulation. In *Robotics: Science and Systems (RSS)*, 2025a.

Yanzhe Wang, Wei Yu, Hao Wu, Haotian Guo, and Huixu Dong. SA-DEM: Dexterous extrinsic robotic manipulation of non-graspable objects via stiffness-aware dual-stage reinforcement learning. *IEEE Transactions on Automation Science and Engineering*, 23:347–362, 2026.

Yuanfei Wang, Xiaojie Zhang, Ruihai Wu, Yu Li, Yan Shen, Mingdong Wu, Zhaofeng He, Yizhou Wang, and Hao Dong. AdaManip: Adaptive articulated object manipulation environments and policy learning. In *International Conference on Learning Representations (ICLR)*, 2025b.

Z. Wang, Y. Mo, S. Jin, and W. Yuan. DoorBot: Closed-loop task planning and manipulation for door opening in the wild with haptic feedback. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16058–16064, 2025c.

R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. Guibas, and H. Dong. VAT-Mart: Learning visual action trajectory proposals for manipulating 3D articulated objects. In *International Conference on Learning Representations (ICLR)*, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yongqiang Yu, Ran Shi, and Yunjiang Lou. Bias estimation and gravity compensation for wrist-mounted force/torque sensor. *IEEE Sensors Journal*, 22(18):17625–17634, 2021.

H. Zhang, B. Eisner, and D. Held. FlowBot++: Learning generalized articulated objects manipulation via articulation projection. In *Conference on Robot Learning (CoRL)*, 2023.

X. Zhang, Y. Wang, R. Wu, K. Xu, Y. Li, L. Xiang, H. Dong, and Z. He. Adaptive articulated object manipulation on the fly with foundation model reasoning and part grounding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13032–13042, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, 2023.

Z. Zhao, Y. Li, W. Li, Z. Qi, L. Ruan, Y. Zhu, and K. Althoefer. Tac-Man: Tactile-informed prior-free manipulation of articulated objects. *IEEE Transactions on Robotics*, 41:538–557, 2025.

A LLM USAGE STATEMENT

We used large language models for proofreading and correcting grammatical errors, assisting with translation, and refactoring the implementation code. All other aspects of this work, including the research ideas, methodology, experiments, and analysis, were carried out entirely by the authors. All LLM-assisted text and code were carefully reviewed and verified by the authors, and we take full responsibility for the final content of this paper.

B LIMITATIONS

BARS-RL trains one policy per object, so a new object requires training a new policy, and learning a single policy across objects remains future work. In addition, the policy has no visual input: it relies on the object state, which we obtained from fiducial markers in our real-world experiments, and on force feedback. Deploying it outside such an instrumented setup therefore requires estimating the object state by other means and closing the sim-to-real gap in force sensing and contact dynamics.

C ADDITIONAL RESULTS

C.1 TRAINING DETAILS WITHOUT CURRICULUM

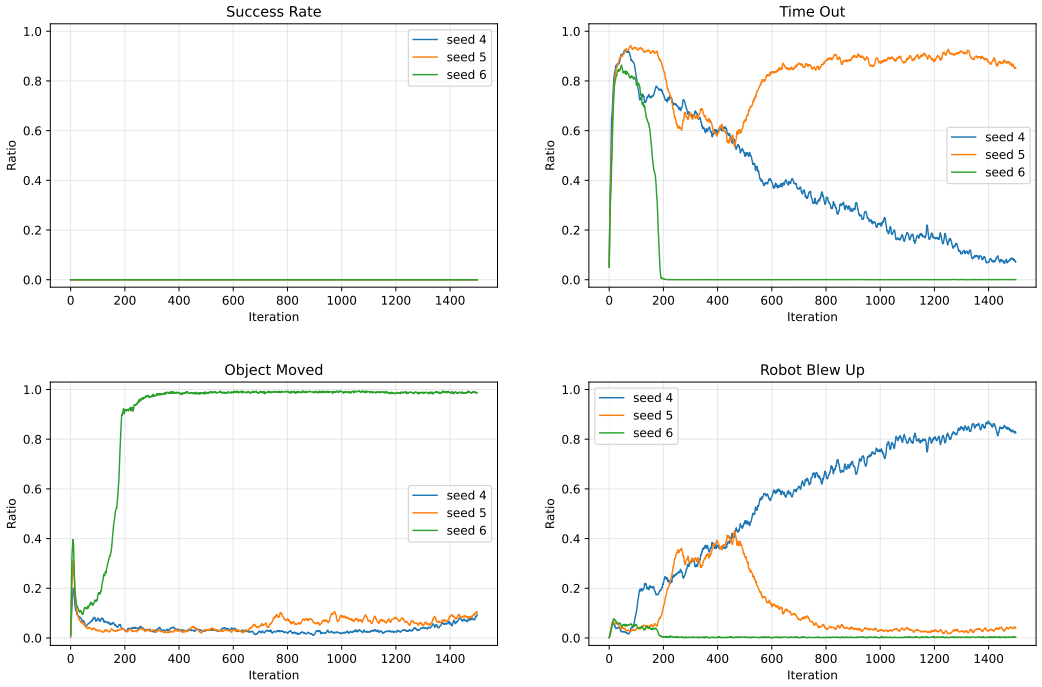


Figure 8: **Episode termination breakdown of BARS-RL w/o curriculum during training**

Figure 8 shows the termination breakdown of BARS-RL w/o curriculum on the mini fridge. Every episode ends in one of four ways: success, time-out, object-body displacement beyond the tolerance, or simulation divergence. We plot three seeds illustrating distinct failure modes. Success remains at zero across all plotted seeds, so the policy never receives the success bonus. In seed 6 (green), nearly every episode terminates due to excessive body displacement within 200 iterations. In seed 5 (orange), most episodes time out. In seed 4 (blue), an increasing fraction of episodes terminates in simulation divergence: since every step incurs the time penalty and the policy never obtains the success bonus, ending an episode early yields a higher return than continuing it, and the policy learns to terminate early.

These observations are consistent with difficulty learning coordinated pulling and stabilization directly from the final task distribution. Our curriculum addresses the coordination difficulty by first establishing opening skills on anchored objects, then introducing unanchored objects that base friction can hold, before training on support-dependent conditions (Section 4.3). With this staged training, BARS-RL achieves the success rates in Table 1, whereas training without the curriculum achieves 0% under the tested setup.

C.2 Applying BARS-RL to New Objects

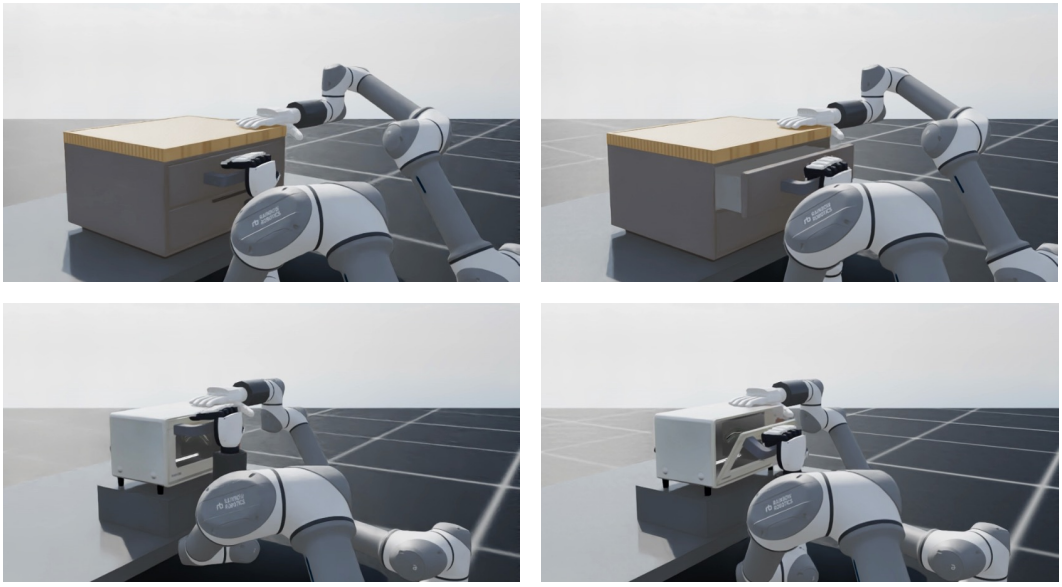


Figure 9: **Applying BARS-RL to new objects.** Top: cabinet with a prismatic drawer. Bottom: oven with a revolute door.

Figure 9 shows rollouts of BARS-RL on the two objects from ArtVIP (Jin et al., 2026) evaluated in Section 5.3. We attached a handle to each object so that the task arm could grasp it, and equipped each closure with the opening-resistance model of Section 4.1.

C.3 Sensitivity to the Base-Slip Cost Parameters

Table 6: **Sensitivity to the base-slip cost parameters.** Success rate (%) on the mini fridge over three seeds, with $\rho \in [1.25, 2.0]$.

Setting	Success rate
Default ($\lambda_{\text{base}} = 0.04$, $\kappa = 0.7$)	$62.2_{\pm 6.1}$
$\lambda_{\text{base}} \times 0.5$	$59.4_{\pm 14.6}$
$\lambda_{\text{base}} \times 2$	$57.2_{\pm 11.2}$
$\kappa = 0.5$	$69.0_{\pm 4.5}$
$\kappa = 0.8$	$62.6_{\pm 3.6}$

To examine the sensitivity to the handcrafted parameters of the reward, we retrained BARS-RL with a different weight λ_{base} or safety factor κ of the base-slip cost. We used the mini fridge, the object with the lowest success rate and the largest drop when c_{base} is removed (Table 4), and thus the object we expected to be most sensitive to these parameters. We changed one parameter at a time and kept all other settings as in Appendix D. Each setting was trained with three seeds, so the default value differs from that of Table 1 (68.4%), which uses five seeds.

Table 6 reports the results. When λ_{base} is halved or doubled, or κ is varied between 0.5 and 0.8, the success rate stays between 57.2% and 69.0%. Three of the four settings differ from the default by at most 5.0 points, which is within its standard deviation across seeds (6.1). The remaining setting, $\kappa = 0.5$, is 6.8 points higher than the default, so the default is not the best-performing setting. Changing λ_{base} in either direction does, however, increase the standard deviation across seeds.

C.4 Sensitivity to Gradual Closure Release

To assess whether BARS-RL depends on an abrupt breakaway assumption, we evaluated the trained policies without retraining under gradual-release variants. Once latch release begins, the remaining opening resistance decreases linearly to zero over a handle displacement of D, measured from the onset of release. Thus, $D = 0$ recovers the original discrete breakaway model, whereas $D = 10$ and 30 mm retain a gradually decreasing resistance over the corresponding travel distances.

For each object and each D, we used the same object model, initial poses, and randomization ranges, and evaluated three policy seeds for 1,024 episodes per seed, totaling 3,072 episodes. We report the mean and standard deviation of success rates across the three seeds for $\rho \in [1.25, 2.0]$. The $D = 0$ results are from this separate three-seed evaluation rather than the five-seed evaluation reported in Table 1.

Table 7: **Sensitivity to gradual closure release.** Success rate (%) for $\rho \in [1.25, 2.0]$ over three seeds.

Release distance D	Microwave	Mini fridge	Air fryer
0 mm (original model)	85.3±2.2	63.2±5.3	95.7±1.1
10 mm	83.9±1.3	62.4±4.1	95.8±1.3
30 mm	80.8±1.7	57.3±4.5	95.3±2.3

Relative to $D = 0$, success changes by at most 1.4 percentage points at $D = 10$ mm and decreases by at most 5.9 points at $D = 30$ mm (Table 7). These results show that the policies remain effective under gradual release without retraining, with only a moderate decrease at $D = 30$ mm.

D Implementation Details

D.1 Episode Reset

The initial pose of an episode is not learned. This section states the rules that define a valid initial pose (Sections D.1.1 and D.1.2) and shows when such a pose exists and how to construct it (Section D.1.3). The same rules and parameter values apply to every object and robot. The initial poses of our experiments were obtained with these rules. At reset, the fingers of the task hand close on the handle in a scripted warm-up before the policy takes control, and the support hand rests on the top surface without pressing.

Inputs. The robot model (kinematics, joint ranges, torque limits $\tau^{\max}$, link masses, hand geometry), the object model (top surface, handle mesh, articulation, mass m, base friction μ_b), and the force limits of the two arms, that is, the largest pulling force and the largest pressing force, $F_{\text{pull}}^{\max}$ and $F_s^{n,\max}$. The breakaway force F_{res} is not used, since it is unobserved; the initial pose is required to remain valid for any pull up to $F_{\text{pull}}^{\max}$.

Arm feasibility. A joint configuration $\mathbf{q} \in \mathbb{R}^n$ *realizes* a hand pose $\mathbf{T}$ under a hand load $\mathbf{F}$ if

$$\begin{aligned} &\mathrm{FK}(\mathbf{q}) = \mathbf{T}, \qquad q_i^{\min} < q_i < q_i^{\max}, \qquad \sigma_{\min}\big(\mathbf{J}(\mathbf{q})\big) > \sigma_0, \\ &\big|\big[\mathbf{J}(\mathbf{q})^\top \mathbf{F} + \mathbf{g}(\mathbf{q})\big]_i\big| < \eta\, \tau_i^{\max}, \qquad d(\mathbf{q}) > d_{\min}. \end{aligned} \tag{10}$$

Here $\mathbf{J}$ is the $6 \times n$ Jacobian at the hand, $\sigma_{\min}$ its smallest singular value (Klein & Blaho, 1987), $\mathbf{g}$ the gravity torque, and d the smallest distance from the arm and hand to the arm bases, the table, the object and the other arm, intended contacts excluded. The torque condition must hold for $\mathbf{F}$ and for $\mathbf{F} = \mathbf{0}$; the torque is affine in the load, so it then holds for every load in between. The

load is $F_{\text{pull}}^{\max}$ along the opening direction for the task arm, and $F_s^{n,\max}$ downward plus the friction $\mu_p F_s^{n,\max}$ under the support hand for the support arm, where μ_p is the friction coefficient between the support hand and the object body.

Parameters. We use $d_{\min} = 20\,\text{mm}$, $\sigma_0 = 0.05$ (m and rad units), $\eta = 0.85$ for the task arm and 0.9 for the support arm, and $\mu_p = 0.5$. These values were chosen empirically: $d_{\min}$ absorbs positioning error, and the same value is used for every clearance below; σ_0 keeps the arm away from configurations where force control becomes unreliable; and η keeps the joint torques from saturating. Each parameter is a one-sided margin, and the results of Section D.1.3 hold for any values. We set the force limits $F_{\text{pull}}^{\max}$ and $F_s^{n,\max}$ below the forces that the real arms can exert; for the support arm, $F_s^{n,\max} = 40\,\text{N}$.

D.1.1 Support-hand rule

The support hand presses the top surface of the object body with its four fingers flat. The rule selects the hand position $\mathbf{c} = (x, y)$ on the surface and its yaw ψ by exhaustive search on a $5\,\text{mm} \times 5^\circ$ grid.

Step 1 (contact). Keep the poses for which (i) the four finger pads and a border of width $d_{\min}$ lie on the top surface of the object body, where the surface slope is below 15° (less than the friction angle $\arctan \mu_p = 26.6^\circ$) and the pad heights differ by less than $5\,\text{mm}$; (ii) the thumb and palm, which protrude below the pad plane, are over free space; (iii) the hand is at least $d_{\min}$ from the volume swept by the moving part; and (iv) the hand is at least $100\,\text{mm}$ from the handle, which leaves room for the task hand.

Step 2 (ranking). The pressing force itself is commanded by the policy, which is trained over a randomized range of ρ (Section 4.3). The rule only ranks the poses by the force they would need. Let $F_{\text{slip}} = \mu_b mg$ as in Section 3, $\rho_{\max} = F_{\text{pull}}^{\max}/F_{\text{slip}}$ the largest resistance-to-friction ratio the initial pose must support, and $f = F_s^n/(mg)$. With the ellipsoidal limit-surface model of planar sliding (Goyal et al., 1991; Howe & Cutkosky, 1996), the object neither slides nor rotates under a pull $F_{\text{pull}}^{\max}$ if

$$\left(\frac{\max(0,\ \rho_{\max}\mu_b - \mu_p f)}{\mu_b(1+f)}\right)^2 \left[1 + \left(\frac{e(\mathbf{c}, f)}{r_c}\right)^2\right] \le 1, \qquad e(\mathbf{c}, f) = \frac{s_g + f\, s_c(\mathbf{c})}{1+f}, \tag{11}$$

where s_g and s_c are the lateral offsets of the center of mass and of the pressing point from the line of the pull, e is the resulting offset of the center of pressure, and r_c is the mean radius of the base contact. For $e = 0$ and $\mu_p = 0$, equation 11 reduces to the bound of Eq. (3), $f \ge \rho_{\max} - 1$. The required pressing force $f^\star(\mathbf{c}, \psi)$ is the smallest f satisfying equation 11.

Step 3 (selection). Sort the poses by $f^\star$, ties broken by distance to the handle, and take the first one that some $\mathbf{q}_{\text{sup}}$ realizes in the sense of equation 10. The distance d includes the task arm at its initial pose, so the two arms are separated by at least $d_{\min}$.

D.1.2 Task-hand rule

Handle frame. The frame is that of the minimum-volume bounding box of the handle mesh, with $\hat{\mathbf{z}}$ along the longest edge (the bar axis) and $\hat{\mathbf{x}}$ along the second longest, pointing away from the body. One reference grasp $\mathbf{T}_{\text{ref}}^{(k)}$, expressed in this frame, is stored per handle type k. The handle is classified by its box dimensions, and a handle of unknown type is rejected.

Step 1 (candidates). Candidates vary the reference grasp by a rotation ϑ about the bar, a shift λ along it, and a change v of the distance to it:

$$\mathbf{T}_{\text{cand}}(\vartheta, \lambda, v) = \mathbf{T}_{\text{handle}}\, \text{Rot}_{\hat{\mathbf{z}}}(\vartheta)\, \text{Trans}\big(\lambda\hat{\mathbf{z}} + v\,\hat{\mathbf{n}}_\perp\big)\, \mathbf{T}_{\text{ref}}^{(k)}, \tag{12}$$

with $|\vartheta| \le \vartheta_{\max} = 40^\circ$, $\lambda \in [-20, 40]\,\text{mm}$ and $v \in [-10, 10]\,\text{mm}$, where $\hat{\mathbf{n}}_\perp$ is the direction from the bar to the hand, normal to the bar.

Step 2 (contact). Keep the candidates whose fingers reach the handle without penetrating it, whose hand clears the table, and whose fingers wrap the bar when closed in simulation, that is, all four distal joints flex by at least $1.2\,\text{rad}$ and the fingers stay on the bar under a pull.

Step 3 (selection). Keep the candidates that some $\mathbf{q}_{\text{task}}$ realizes in the sense of equation 10 and for which the handle remains reachable while the object opens (condition (iii) of Section D.1.3). Among them take the one with the largest torque margin $\min_i(1 - |\tau_i|/\tau_i^{\max})$. The support pose and the grasp are coupled only through the distance between the arms; if one choice leaves no candidate for the other hand, the next one in order is taken.

D.1.3 Existence of an initial pose

Whether the rules have a solution depends on where the object stands relative to the robot. We give a necessary and sufficient condition for a placement that both hands can reach, a closed-form construction of it, and show that the valid placements form a set of positive area.

Setting. Both arms are mounted upright on a horizontal plane, so that their first-joint axes are vertical, a distance $b > 0$ apart. Subscripts t and s denote the task and support arms. The object stands on a horizontal surface, so its placement is $\mathbf{x} = (p_x, p_y, \beta) \in SE(2)$. Fix one support pose and one grasp candidate that satisfy the contact conditions of Sections D.1.1 and D.1.2, which do not depend on the placement. The hand poses $\mathbf{T}_s$ and $\mathbf{T}_t$ at reset are fixed in the object. For a hand, let $r > 0$ be the horizontal distance of its reference point from the first-joint axis of its arm, and φ its yaw angle measured from the radial direction. Independently of $\mathbf{x}$, the object fixes the height and tilt of both hands, the horizontal distance $d_h > 0$ between their reference points, and their yaw angles α_t, α_s measured from the line joining these points.

We call the conditions of equation 10 other than the range of the first joint and the distance *d intrinsic*.

Proposition 1. *For a given hand height, tilt and load fixed in the hand frame, whether a hand pose has a configuration satisfying the intrinsic conditions depends only on* (r, φ).

Proof. Two poses with equal (r, φ), height and tilt differ by a rotation about the first-joint axis, which is produced by shifting the first joint alone. The other joints are unchanged; $\mathbf{J}$ is only rotated, so its singular values are unchanged; the axis is vertical, so the gravity torque is unchanged; and the load rotates with the hand, so $\mathbf{J}^\top \mathbf{F}$ is unchanged. □

The *reachable set* $\mathcal{F}_a$ of arm $a \in \{t, s\}$ is the set of such (r, φ). It is computed once on a grid, with one configuration stored per grid point.

Theorem 1. *A placement at which both arms satisfy the intrinsic conditions exists if and only if there are* $(r_t, \omega_t + \alpha_t) \in \mathcal{F}_t$ *and* $(r_s, \omega_s + \alpha_s) \in \mathcal{F}_s$ *such that*

$$\ell(r_t, \omega_t, r_s, \omega_s) := \left| d_h + r_t e^{-i\omega_t} - r_s e^{-i\omega_s} \right| = b. \tag{13}$$

In particular, if $\mathcal{F}_t \times \mathcal{F}_s$ *is connected, such a placement exists whenever* $\inf \ell < b < \sup \ell$.

Proof. Identify the plane with $\mathbb{C}$, with arm bases B_t and $B_s = B_t + b$ and hand points $P_a = B_a + r_a e^{i\theta_a}$. The object imposes $P_s - P_t = d_h e^{i\gamma}$ and the hand yaw angles $\gamma + \alpha_a$, so $\varphi_a = \gamma + \alpha_a - \theta_a$. With $\omega_a = \varphi_a - \alpha_a$ we have $\theta_a = \gamma - \omega_a$ and

$$b = e^{i\gamma}\big(d_h + r_t e^{-i\omega_t} - r_s e^{-i\omega_s}\big). \tag{14}$$

Taking moduli gives necessity. Conversely, given a solution of equation 13, the bracket is nonzero; let γ be minus its argument and $\theta_a = \gamma - \omega_a$. Then equation 14 holds, so the unique planar placement that puts the two hand points at P_t and P_s gives the hands the coordinates $(r_a, \omega_a + \alpha_a) \in \mathcal{F}_a$. The last claim holds because a continuous function maps a connected set onto an interval. □

Closed-form construction. Choose $(r_t, \omega_t + \alpha_t) \in \mathcal{F}_t$ and r_s, and let $w = d_h + r_t e^{-i\omega_t}$. If $w \neq 0$ and $\big||w| - r_s\big| \le b \le |w| + r_s$, equation 13 is solved by

$$\omega_s = -\arg w \pm \arccos \frac{|w|^2 + r_s^2 - b^2}{2|w| r_s}, \qquad \gamma = -\arg\big(w - r_s e^{-i\omega_s}\big). \tag{15}$$

If $(r_s, \omega_s + \alpha_s) \in \mathcal{F}_s$, the placement follows from $\theta_a = \gamma - \omega_a$, and the joint configurations are the stored ones with the first joint shifted (Proposition 1). No inverse kinematics is solved again.

Placement region. A placement $\mathbf{x}$ is *valid* for the fixed pair if

(i) some $\mathbf{q}_{\mathrm{sup}}$ realizes $\mathbf{T}_s$;

(ii) some $\mathbf{q}_{\mathrm{task}}$ realizes $\mathbf{T}_t$;

(iii) the handle remains reachable: for every articulation state q between the closed state q_0 and the largest target q_g there are $|\vartheta| \le \vartheta_{\max}$ and a configuration of the task arm, within the joint ranges, with $\sigma_{\min}(\mathbf{J}) > 0$ and gravity torque below $\eta\tau^{\max}$, whose hand holds the handle at q with the grasp rotated by ϑ about the bar.

Condition (iii) does not prescribe how the policy moves. It only excludes placements where the handle leaves the workspace of the task arm, so that opening would be impossible for any policy that keeps the hand on the handle. The margin σ_0 is required only at reset, where the breakaway force must be produced. The *placement region* $\mathcal{W} \subset SE(2)$ is the union of the valid placements over the finitely many pairs that the rules test.

Proposition 2. *$\mathcal{W}$ is open. Hence, if $\mathcal{W}$ is nonempty, every placement in it tolerates a position error and a yaw error of some positive size.*

Proof. It suffices to treat one pair. Every condition is a strict inequality whose left-hand side is continuous in the configuration and the placement, and holds at a configuration where $\mathbf{J}$ has rank 6. By the implicit function theorem (for $n > 6$, with the redundant joints held fixed), the configurations of (i) and (ii) at $\mathbf{x}_0$ extend continuously to placements near $\mathbf{x}_0$, where the inequalities persist. For (iii), fix $q' \in [q_0, q_g]$ with its ϑ and configuration. The target hand pose is continuous in $(q, \mathbf{x})$, so the same argument gives a neighborhood of $(q', \mathbf{x}_0)$ on which (iii) holds with the same ϑ. Finitely many of these neighborhoods cover the compact set $[q_0, q_g] \times \{\mathbf{x}_0\}$, and the intersection of their $\mathbf{x}$-parts is a neighborhood of $\mathbf{x}_0$ on which (iii) holds for all q. □

Summary. For a given object and robot, an initial pose exists if (H1) $\mathcal{F}_t$ and $\mathcal{F}_s$ are nonempty at the two hand heights; (H2) equation 13 has a solution; and (H3) at one of the placements given by equation 15, the first joint of both arms is within range, $d > d_{\min}$, and (iii) holds. If (H1) or (H2) fails, no initial pose exists. (H1) and (H2) are decided from $\mathcal{F}_t$ and $\mathcal{F}_s$ in closed form, and (H3) by a direct test of the constructed placements. When (H3) holds, $\mathcal{W}$ has positive area, and a mobile manipulator only needs to bring the object into $\mathcal{W}$.

Our setting. For the simulated robot (two RB5-850E arms, $b = 1.0\,\mathrm{m}$) and each of the five objects of Section 5 and Appendix C.2, (H1) and (H2) hold, and the first-joint ranges and condition (iii) hold at placements constructed with equation 15; the distance condition is tested by the rules when the initial pose is selected. This remains true when σ_0 or μ_p is halved or doubled, and when η is set to $(0.70,\ 0.75)$ or $(1,\ 1)$.

D.1.4 Limitations

The rules handle bar-shaped handles only; recessed handles, knobs and levers are rejected at classification. The task hand is assumed to hold the handle and to stay near its initial grasp, within a rotation $\vartheta_{\max}$ about the bar. The support hand always presses with its fingers horizontal; tilted pressing poses are not considered.

D.2 Observation and Action Spaces

The actor observes only quantities that are measurable or estimable on the real robot (Table 8). The object component of the observation is the 10-dimensional vector

$$o_t^{\mathrm{obj}} = [\, q_t,\ q_g,\ q_g - q_t,\ \Delta p_o,\ \Delta\theta_o,\ v_o \,], \tag{16}$$

where $\Delta p_o \in \mathbb{R}^3$ is the translation of the object body since reset, $\Delta\theta_o \in \mathbb{R}$ is its rotation since reset, and $v_o \in \mathbb{R}^3$ is its translational velocity. The critic receives the actor observation together with the physical parameters $\zeta = (F_{\mathrm{res}}, \mu_{\mathrm{b}}, m)$, the base-slip margin δ_{b} of equation 1, and the base-contact force. These quantities are used only during training and never enter the actor.

Table 8: Actor observation.

Component	Dim.
Arm and hand proprioception (both arms)	48
End-effector poses (both arms)	14
Six-axis wrist F/T (both arms)	12
Per-pad tactile signals (both hands)	104
Previous action	25
Object vector, equation 16	10
Total	213

The policy outputs a 25-dimensional action at 30 Hz (Table 9). Each arm receives a 12-dimensional operational-space command consisting of a relative end-effector pose target and a wrench target, and the hand of the task arm receives one additional grip command. A fixed hybrid force/position selection matrix assigns one translational axis to force control and the remaining five Cartesian directions to position control. For the task arm, the force-controlled axis follows the opening direction $\hat{\mathbf{a}}(q_t)$, and for the support arm, it follows the normal of the support hand contact surface. A low-level operational-space controller converts the policy commands into joint-torque commands at 120 Hz.

Table 9: Policy action.

Component	Dim.
Task arm relative pose target	6
Task arm wrench target	6
Support arm relative pose target	6
Support arm wrench target	6
Task hand grip command	1
Total	25

D.3 REWARD

This section gives the closed forms, weights, and thresholds of the reward in equation 8.

Task reward. The progress term is the normalized reduction in articulation error, credited only while the task arm holds the handle,

$$r_{\text{prog}} = \mathbf{1}_{\text{grasp}} \frac{|q_g - q_{t-1}| - |q_g - q_t|}{|q_g - q_0|}, \tag{17}$$

where $\mathbf{1}_{\text{grasp}}$ is active while the task arm maintains its grasp on the handle, which we detect from the fingertip contact sensors: at least one finger reports a contact force above $1\,\text{N}$ (debounced over three control steps, $100\,\text{ms}$, to bridge single-frame contact dropouts) and the nearest fingertip lies within $40\,\text{mm}$ of the handle axis. The success bonus r_{succ} is awarded once, at the step when

$$\mathbf{1}_{\text{succ}} = \mathbf{1}[\,|q_g - q_t| \le \epsilon_q\,]\cdot\mathbf{1}[\,|\dot{q}_t| \le \epsilon_{\dot{q}}\,]\cdot\mathbf{1}_{\text{grasp}}\cdot\mathbf{1}\left[\max_{t' \le t} \|\Delta p_o(t')\| \le \epsilon_p\right]\cdot\mathbf{1}[\,|\Delta\theta_o| \le \epsilon_\theta\,]\cdot\mathbf{1}[\,\delta_{\text{b}} \ge 0\,], \tag{18}$$

and the episode terminates at that step. No hold duration is required; the velocity clause rules out a moving part that coasts through the goal, the displacement clause uses the peak displacement of the object body over the episode, the rotation clause bounds its rotation, and the last clause requires the base-slip margin of equation 1 to be non-negative. For success, $\mathbf{1}_{\text{grasp}}$ uses the stricter, non-debounced contact test (all four fingers loaded on the handle for the microwave and air fryer; thumb plus at least one finger for the mini fridge). The pull term rewards force applied along the opening direction before breakaway, the settling term rewards small articulation error together with low articulation velocity, and the action cost penalizes changes between consecutive actions,

$$r_{\text{pull}} = \mathbf{1}_{\text{grasp}} \cdot \mathbf{1}[\,q_t \text{ short of } q_g\,] \cdot \tanh\left(\frac{\max(F_{\text{pull}}, 0)}{\alpha_{\text{sat}} F_{\text{slip}}}\right), \tag{19}$$

$$r_{\text{settle}} = \mathbf{1}_{\text{grasp}} \cdot \exp\left(-\frac{(q_g - q_t)^2}{\sigma_q^2}\right) \exp\left(-\frac{\dot{q}_t^2}{\sigma_{\dot{q}}^2}\right), \tag{20}$$

$$c_{\text{act}} = \frac{1}{|\mathcal{A}|} \|a_t - a_{t-1}\|_2^2, \tag{21}$$

where F_{pull} is the pulling force of equation 4, $F_{\text{slip}} = \mu_{\text{b}}\, m\, g$ is the per-episode slip threshold of Section 3, and $|\mathcal{A}|$ is the action dimension.

Stabilization terms. The base-slip cost c_{base} is defined in equation 6 with safety factor κ, and the support reward r_{sup} in equation 7 with length scale σ_o. The indicators in equation 7 compare the applied forces with fixed thresholds,

$$\mathbf{1}_{\text{pulling}} = \mathbf{1}_{\text{grasp}} \cdot \mathbf{1}\left[F_{\text{pull}} \ge \bar{F}_{\text{pull}} \;\vee\; \dot{q}_t \ge \bar{\dot{q}}\right], \qquad \mathbf{1}_{\text{pressing}} = \mathbf{1}\left[F_{\text{s}}^n \ge \bar{F}_{\text{press}}\right], \tag{22}$$

where F_{s}^n is the normal force of the support arm in equation 2. The pulling-force threshold is a fraction of the per-episode slip threshold, $\bar{F}_{\text{pull}} = \alpha_{\text{pull}} F_{\text{slip}}$, and the velocity clause keeps the indicator active through the release of the closure, when the moving part starts to move and the wrist load drops to joint friction. The displacement cost penalizes translation and rotation of the object body, the overgrip cost penalizes grip force beyond what the handle load requires, and the off-axis cost penalizes wrist force that does not contribute to opening,

$$c_{\text{body}} = \frac{\|\Delta p_o\|}{\sigma_p} + \frac{|\Delta\theta_o|}{\sigma_\theta}, \tag{23}$$

$$c_{\text{overgrip}} = \frac{1}{F_{\text{grip}}^{\max}} \max\left(F_{\text{grip}} - \frac{F_{t,\text{handle}}}{\mu_{\text{tip}}} - \delta_{\text{safe}},\, 0\right), \tag{24}$$

$$c_{\text{offaxis}} = \frac{\|\bar{\mathbf{f}}_{\text{w}}(t) - F_{\text{pull}}\, \hat{\mathbf{a}}(q_t)\|}{\sigma_{\text{off}}}, \tag{25}$$

where F_{grip} and $F_{t,\text{handle}}$ are the summed normal and tangential contact forces between the fingertips of the task-arm hand and the handle, so $F_{t,\text{handle}}/\mu_{\text{tip}}$ is the grip the current handle load requires and δ_{safe} is a free band above it; $\bar{\mathbf{f}}_{\text{w}}(t)$ and $\hat{\mathbf{a}}(q_t)$ are defined in equation 4. Each cost is clipped at its full scale (c_{body} at 2, c_{act} at 4, the rest at 1), which removes the tail of diverged simulation states without changing the terms inside the operating range.

Weights and parameters. Table 10 lists the weights of equation 8; the stage-dependent weights w_{stab} and w_{sup} are given in Appendix D.5. Table 11 lists the remaining parameters.

Table 10: Reward weights.

Weight	Applies to	Value
λ_{prog}	r_{prog}	2.0
λ_{succ}	r_{succ}	10
λ_{pull}	r_{pull}	0.02
λ_{settle}	r_{settle}	0.02
λ_{sup}	r_{sup}	0.02
λ_{time}	per step	0.003
λ_{act}	c_{act}	0.01
λ_{body}	c_{body}	0.02
λ_{base}	c_{base}	0.04
$\lambda_{\text{overgrip}}$	c_{overgrip}	0.04
λ_{offaxis}	c_{offaxis}	0.04

Table 11: Reward parameters. Angle-based values apply to the revolute doors.

Symbol	Description	Value
κ	Safety factor of the base-slip cost, equation 6	0.7
σ_o	Length scale of the support reward, equation 7	30 mm
α_{pull}	Pulling-force threshold $\bar{F}_{\text{pull}} = \alpha_{\text{pull}} F_{\text{slip}}$, equation 22	0.4
$\bar{\dot{q}}$	Articulation-velocity clause of $\mathbf{1}_{\text{pulling}}$, equation 22	0.1 rad/s
$\bar{F}_{\text{press}}$	Support normal-force threshold for $\mathbf{1}_{\text{pressing}}$, equation 22	5 N
ϵ_q	Articulation tolerance for success, equation 18	5°
$\epsilon_{\dot{q}}$	Articulation velocity tolerance for success, equation 18	0.1 rad/s
ϵ_p	Object-body displacement tolerance for success, equation 18	30 mm
ϵ_θ	Object-body rotation tolerance for success, equation 18	5°
α_{sat}	Saturation of r_{pull} in units of F_{slip}, equation 19	1.4
σ_q	Angle scale of r_{settle} ($= 2\epsilon_q$), equation 20	10°
$\sigma_{\dot{q}}$	Velocity scale of r_{settle} ($= 4\epsilon_{\dot{q}}$), equation 20	0.4 rad/s
σ_p	Translation scale of c_{body}, equation 23	50 mm
σ_θ	Rotation scale of c_{body}, equation 23	5°
$F_{\text{grip}}^{\max}$	Grip-force normalization of c_{overgrip}, equation 24	50 N
δ_{safe}	Free grip band above the required grip, equation 24	5 N
σ_{off}	Force scale of c_{offaxis}, equation 25	30 N

D.4 TRAINING

We optimize the actor and the critic of the asymmetric actor–critic with Proximal Policy Optimization (PPO) (Schulman et al., 2017). Both networks are multilayer perceptrons. The actor receives the observation of Appendix D.2, and the critic additionally receives the privileged inputs listed there (Pinto et al., 2018). Table 12 lists the training hyperparameters.

Table 12: PPO hyperparameters.

Hyperparameter	Value
Parallel environments	2048
Rollout length (steps per environment)	32
Epochs per update	5
Minibatch size	16384
Learning rate	1e-3
Optimizer	Adam (Kingma & Ba, 2015)
Activation	ELU (Clevert et al., 2016)
Discount γ	0.99
GAE λ	0.95
Clip ratio ϵ	0.2
Entropy coefficient	0.005
Value loss coefficient	1.0
Actor MLP hidden sizes	[512, 256, 128]
Critic MLP hidden sizes	[512, 256, 128]

D.5 CURRICULUM

Table 13 lists the settings of each curriculum stage in Algorithm 1. Observation noise (Table 14) is enabled only in the final phase.

Table 13: Curriculum stages. α_s is the probability that the object body is anchored, $\rho \sim \mathcal{U}(\mathcal{R}_s)$, and τ_s is the success threshold for advancing to the next stage.

Phase	Stage	α_s	$\mathcal{R}_s$	w_{stab}	w_{sup}	Obs. noise	τ_s
0	0	1	$\{0\}$	0	0	off	0.7
1	1	0.75	$[0.1, 0.5]$	0.5	1	off	0.3
1	2	0.50	$[0.1, 0.5]$	0.5	1	off	0.5
1	3	0.25	$[0.1, 0.5]$	0.5	1	off	0.7
1	4	0	$[0.1, 0.5]$	0.5	1	off	0.8
2	5	0	$[0.4, 2.0]$	1	1	on	none

D.6 Domain Randomization

Table 14 lists the randomized parameters in the final phase.

Table 14: Randomization parameters.

Group	Parameter	Distribution	Operation	Sampled
Object	Mass m	$\mathcal{U}(0.85, 1.15)$	Scaling	per env
Object	Base friction μ_{b}	$\mathcal{U}(0.15, \mu_{\mathrm{b}}^{\max})$	Absolute	per env
Object	Breakaway force F_{res}	$\rho\, F_{\mathrm{slip}}$, $\rho \sim \mathcal{U}(0.4, 2.0)$	Sampling	per episode
Object	Goal q_g (revolute)	$\mathcal{U}(25^\circ, 80^\circ)$	Sampling	per episode
Object	Goal q_g (prismatic)	$\mathcal{U}(0.28, 0.89)$ of stroke	Sampling	per episode
Robot	Fingertip friction	$\mathcal{U}(0.8, 1.2)$	Scaling	per env
Robot	Support hand pad friction	$\mathcal{U}(0.8, 1.2)$	Scaling	per env
Observation	Wrist F/T	$\mathcal{U}(-1.0, 1.0)$ N	Additive	per step
Observation	Tactile (task arm)	$\mathcal{U}(-0.2, 0.2)$	Additive	per step
Observation	Tactile (support arm)	$\mathcal{U}(-0.5, 0.5)$	Additive	per step

D.7 Baseline Implementation

Data. For each object, all imitation learning baselines were trained on the same 100 demonstrations, collected with a scripted expert that observes ζ and the closure state, with ρ sampled from $[0, 2]$.

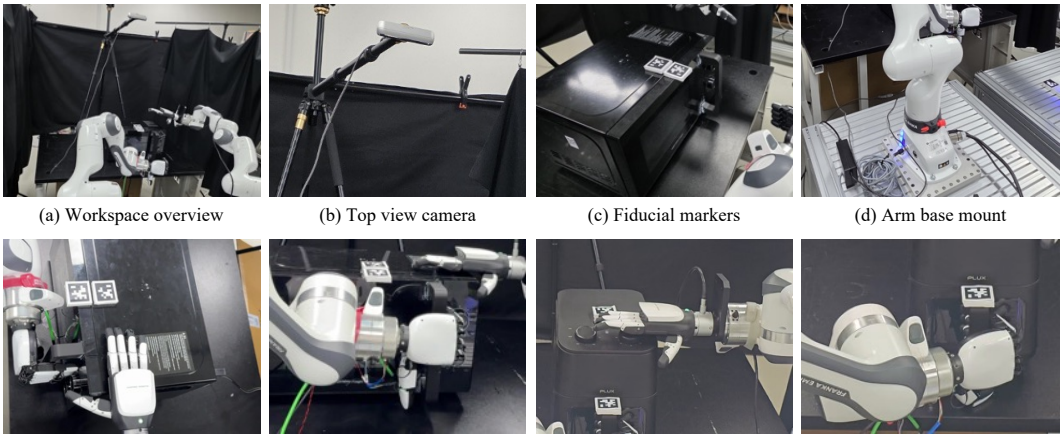
Observations. ACT and Diffusion Policy observe RGB images from a front camera and two wrist cameras, the 24 arm and hand joint positions, and the goal q_g. VoxAct-B observes a voxel grid instead of images: the RGB-D images of the same three cameras are back-projected into a point cloud with the camera calibration, and a fixed workspace is discretized into 50^3 voxels. The F/T variants additionally observe the six-axis wrist F/T; in VoxAct-B, the acting and stabilizing policies each receive the wrist F/T of their own arm. The F/T + tactile variants further observe the per-pad tactile signals that BARS-RL uses (Table 8). None of the imitation learning baselines observes the object-state vector of BARS-RL, and BARS-RL does not observe images. Table 1 therefore compares each method with its own inputs, not different learning algorithms under identical observations.

Actions and control. All three baselines output the same 25-dimensional action as BARS-RL (Table 9) at 30 Hz, and the same 120 Hz hybrid operational-space controller executes it (Appendix D.2).

Adaptation of VoxAct-B. To fit this action and controller, we made three changes to VoxAct-B. First, the voxel grid covers a fixed region of the workspace instead of a crop centered on the object. Second, the keyframe classification head is replaced by a regression head that outputs the continuous action. Third, the motion planner is removed, since the policy issues commands directly to the controller.

Training and evaluation. Each baseline was trained with one seed per object, and every input variant of a baseline used the same settings. ACT was trained for 40,000 steps with a batch size of 32 and a learning rate of 3×10^{-5} (10^{-5} for the backbone), Diffusion Policy for 40,000 steps with a batch size of 64 and a learning rate of 10^{-4}, and VoxAct-B for 10,000 steps with a batch size of 2. All baselines were evaluated at their final checkpoint, without validation-based selection. The two ranges of ρ were evaluated separately: ACT and Diffusion Policy for 1,024 episodes and VoxAct-B for 256 episodes per object and range. Every evaluation episode has a time limit of 10 s (300 control steps). The training loss of VoxAct-B had converged by 10,000 steps. As a check, we trained its variant without F/T input for 50,000 steps: the success rate changed by at most 1.9 points on every object and range of ρ, and its average was 37.5% for $\rho < 1.25$ and 21.7% for $\rho \geq 1.25$, compared with 38.2% and 22.4% at 10,000 steps. The imitation learning rows of Table 1 thus report the success rate of a single training seed, which is why they have no standard deviation.

E REAL-WORLD SETUP



(a) Workspace overview (b) Top view camera (c) Fiducial markers (d) Arm base mount

(e) Microwave: support hand (f) Microwave: handle grasp (g) Air fryer: support hand (h) Air fryer: handle grasp

Figure 10: **Real-world setup.** (a) Workspace with the two arms and the object on the table. (b) Top-view camera. (c) Fiducial markers on the object body and the moving part. (d) Arm base mount. (e–h) Initial contact of the support hand and grasp of the task hand on the microwave and the air fryer.

Figure 10 shows the setup of the real-world experiments in Section 5.2. The task arm is a Franka FR3 with an XHand1 and the support arm is a Franka Panda with an Inspire hand; the two bases are mounted 0.80 m apart. Each object rests on the table without being fixed, at the placement used to compute the initial poses (Appendix D.1). Fiducial markers on the object body and the moving part are tracked by the overhead camera to obtain the object state. For the pull-force measurements of Figure 6, the robot arm pulled each handle at 3 mm/s. Force logs were available for 13 of the real-world trials (10 on the microwave and 3 on the air fryer). In these trials, the pressing force that the policy commanded stayed within 40 N. The pressing force estimated from the joint torques averaged 13–31 N per trial and peaked at 48.4 N, exceeding 40 N for at most 0.33 s, and the 80 N safety stop was never triggered. These estimates are logged at 30 Hz and do not capture millisecond-scale impacts.

F QUALITATIVE ROLLOUTS

Figure 11 shows simulated rollouts of BARS-RL on the three objects.

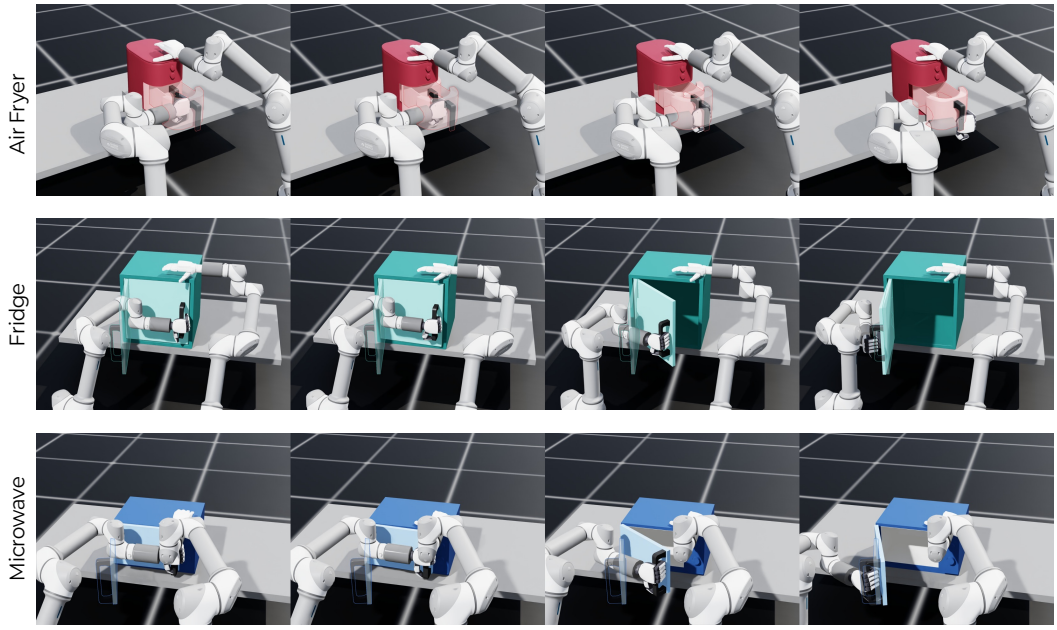


Figure 11: **Rollouts of BARS-RL on the three objects.** Each row shows one object (air fryer, mini fridge, and microwave from top to bottom), with frames ordered from left to right over time.