Research Article

Nazakat: Compliant morphology and developmental learning for robotic manipulation

Pari Sharma*, Vedant Belsare, Rudra Jaiswal, Riddhesh Dalvi, Pragati Khare, Manisha Tiwari

Department of Computer Engineering, Mukesh Patel School of Technology Management & Engineering, SVKM's NMIMS, Mumbai 400056, India.

***Correspondence to:** Dr. Pari Sharma, Department of Computer Engineering, Mukesh Patel School of Technology Management & Engineering, SVKM's NMIMS, Mumbai 400056, India. E-mail: parisharma1606@gmail.com

Received: 07 October 2026 | Approved: 27 October 2026 | Online: 27 October 2026

Abstract

Robotic manipulation of unknown objects remains difficult because a gripper cannot know an object's firmness, shape, or slipperiness in advance. We present Nazakat (Naz), a controlled 2 × 2 factorial simulation study crossing gripper morphology (rigid *vs.* cephalopod-inspired soft) with learning strategy (sparse-reward reinforcement learning (RL) *vs.* a six-stage developmental curriculum modeled on infant motor milestones), trained with Soft Actor-Critic and evaluated on twelve objects across three random seeds per condition. Compliance produced the largest and most consistent effect: the soft gripper improved generalization over the rigid gripper in every seed-matched comparison under both learning strategies, although soft-gripper standard RL was also

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

the least reproducible condition (36.5 ± 42.1% generalization; 10.6% to 85.0% across seeds). The curriculum's early stages taught reaching and grip discovery that standard RL did not, but a later-stage uncapped force-overshoot penalty led policies to avoid sustained grasps: rigid-gripper curriculum runs reached 0% on all three seeds, and soft-gripper curriculum runs had to be stopped at Stage 3 (18.1 ± 10.5% generalization). Softening or ramping this penalty did not yield a reliable fix, pointing to the raw joint-angle action interface as a further bottleneck. These results support a directional, simulator-specific conclusion that compliance helps, and a negative result about this particular curriculum rather than about staged learning in general.

Keywords: Reinforcement learning, soft robotics, developmental robotics, robotic grasping, compliant manipulation, curriculum learning

INTRODUCTION

Robots that function without any physical contact or supervision of human operators in distant environments like a planetary environment, deep sea, or potentially dangerous locations such as collapsed buildings regularly need to work with objects that have not been previously known in terms of their fragility, shape and surface characteristics. Real-time human guidance often isn't an option in these settings, since communication delay alone can rule it out. When that's the case, a robot that runs into something unexpected doesn't fail gracefully. It just fails, and often for good. NASA's Spirit rover is a clear example: in 2009 it drove onto a patch of loose, sandy terrain that had looked perfectly solid from orbital imagery, sank in immediately, and never got free again, ending a mission that had already outlived its planned lifetime by years. Nothing about that failure traced back to a computational shortfall or a broken part. The rover simply had no way to know, before committing to that drive, how that particular patch of ground would actually behave underneath it.

Manipulation faces the same problem in miniature, every time a gripper closes on an object it has not encountered before. Most industrial and research manipulators handle this by narrowing the problem rather than solving it: rigid, servo-actuated grippers achieve high precision and repeatability on objects whose shape, weight, and fragility are known ahead of time, and control policies, whether hand-engineered or learned, are typically developed and evaluated against that same fixed hardware. This is a reasonable simplification for isolating algorithmic contributions, but it leaves an open question that neither side of the literature answers on its own: whether a robust grasp against an unfamiliar object succeeds because the hand adapted physically, because the policy learned to compensate for a hand that could not adapt, or because the two properties interact in a way that neither produces alone.

Two research traditions offer partial answers. Compliant and biomimetic gripper design, most notably grippers inspired by the octopus arm, which has no rigid skeleton and distributes contact force across its surface rather than concentrating it at discrete points, has demonstrated real improvements in fragile-object handling and tolerance to misalignment, without requiring a control policy to compensate explicitly for hardware limitations. Separately, learned control, reinforcement learning (RL) in particular, has produced manipulation policies that reach very high reliability on novel objects, but almost always paired with a single, fixed gripper design, placing the entire adaptive burden on the policy. A third, smaller literature asks whether the order in which a learning system encounters tasks matters as much as the learning algorithm itself, drawing on developmental psychology's observation that human infants build manipulation competence in stages: undirected reaching, accidental contact, accidental grasping, force calibration, dexterity, and finally manipulation while in motion, rather than being trained end-to-end against a single objective from the start.

We did not find a study that varies gripper morphology and learning curriculum independently within one controlled experimental design. Studies that examine

compliant hardware typically pair it with a fixed heuristic or lightly-tuned controller; studies that examine learned control typically pair it with fixed, usually rigid, hardware. This leaves the interaction between the two factors, whether they are additive, redundant, or synergistic, essentially unmeasured.

This paper addresses that gap directly. We present Nazakat (Naz), a controlled 2 ×2 factorial study crossing gripper morphology (rigid *vs.* a cephalopod-inspired soft gripper) with learning strategy (a standard sparse-reward RL baseline *vs.* a six-stage developmental curriculum modeled explicitly on infant motor-development milestones), trained and evaluated on a shared simulated platform under identical conditions. The four resulting experimental conditions, namely rigid + standard (baseline), soft + standard (hardware contribution), rigid + developmental (learning contribution), and soft + developmental (full system), let us isolate what each factor contributes individually and in combination, evaluated across a 12-object benchmark spanning rigid, fragile, deformable, and slippery objects, with the simulation additionally supporting dynamically disturbed-base evaluation as a final curriculum stage (a simulated planetary rover platform used here purely as a demanding, physically grounded stress test, not as a hardware contribution in itself). Our central research question is what contributes more to robust robotic manipulation: compliant physical morphology, biologically inspired learning, or their combination, and whether the two factors add or multiply.

The remainder of this paper is organized as follows. Section 2 reviews the compliant-morphology, RL-for-manipulation, developmental-learning, and sim-to-real literatures in turn and identifies the specific gap this work addresses. Section 3 describes the simulation environment, the four experimental conditions, the RL formulation for each, and the evaluation protocol. Section 4 reports results for each condition, including a multi-seed comparison. Section 5 discusses the findings against our research question, Section 6 states the limitations of this study, Section 7 outlines the planned physical

hardware realization of this system as a separate follow-up study, and Section 8 concludes.

RELATED WORK

Overview

Robust delicate manipulation, the ability to pick up, hold, and release objects of unknown fragility, geometry, and surface friction without prior knowledge of their properties, has been approached from two largely separate research traditions. One treats the problem as a hardware question, building compliant end-effectors whose physical structure absorbs uncertainty a rigid gripper cannot. The other treats it as a control question, using RL to let a policy discover grasp strategies no engineer explicitly programmed. Broader work on autonomous planetary manipulation shows the value of allowing robots to learn about unknown objects through interaction rather than relying entirely on prior models[2]. Reviews of space robotics similarly emphasize the need for greater autonomy when communication delays, environmental uncertainty, and incomplete models limit continuous human supervision[3]. Related work on myoelectric prosthetic hands also shows the importance of combining machine learning with sensing and efficient embedded computation when autonomy must operate under hardware constraints[4]. A third, smaller body of work asks whether the sequence in which a learning system is exposed to tasks matters as much as the algorithm itself, drawing on developmental psychology's argument that competence should be built in stages rather than trained end-to-end against a single objective. What is largely absent is a controlled study that varies hardware and learning strategy independently and asks which factor, or which combination, actually produces the observed performance. This review traces each tradition in turn before identifying that gap precisely.

Compliant and biomimetic gripper morphology

The argument for compliant hardware stems from the straightforward idea that a gripper that physically conforms to anything has far fewer demands for perception and control

than one that does not. This reasoning attends back to very early octopus-inspired versions of constructions, which made use of soft arms and other bioinspired constructions that could elongate, contract, twist, move and manipulate objects only because of the use of the same compliant structure. More recent studies targeted this lineage towards more practical applications, such as sensorized soft arms, sucker-inspired grippers, multi-functional underwater grippers, and a bigger number of review papers discussing the use of sensing and control principles in this area[5-12].

A separate line of work has tried to actually quantify what compliance buys you, and where its limits are. Li *et al.*'s scoping review compared two dozen soft-rigid grippers against industrial parallel rigid designs and found clear trade-offs between compliance, how many object types a gripper can handle, how much force it can apply, and its range of motion[5]. Zhou *et al.* took the discussion further, pairing a soft gripper with learning-based object recognition, force control, and slip compensation in one system[13], and James and Lepora showed something related from the control side: tactile slip detection alone can stabilize a grasp just by increasing grip force the moment an object starts to slip[14]. Put together, these results point to something the literature doesn't always acknowledge outright: morphology and learning aren't as independent in practice as they're usually treated on paper.

Taken together, this body of work makes two things clear. First, compliant and cephalopod-inspired morphology is now an established design pattern with demonstrated benefits for delicate handling, rather than a novel proposal in itself[5-12]. Second, and more importantly for the present study, the literature does not provide a clean comparison between matched rigid and compliant hardware under the same learned policy. The morphology literature therefore establishes that compliance can help, but it does not isolate how much of the benefit is attributable to the body rather than to the control scheme paired with it.

Learning-based control for robotic manipulation

The RL side of the manipulation literature has moved rapidly. Han *et al.*'s survey of deep RL for manipulation maps the field across value-based, policy-based, and actor-critic families and concludes that no single algorithm dominates across all tasks[15]. Newbury *et al.*'s systematic review of 85 grasp-synthesis papers similarly shows that most reported successes rely heavily on simulation, while high-quality real-world grasp datasets remain more limited[16].

Several landmark systems illustrate what learned control can achieve once trained at sufficient scale. Mahler *et al.*'s Dex-Net 4.0 learns a unified policy for parallel-jaw and suction grippers from five million synthetic depth images and reports greater than 95% reliability on bins containing up to 25 novel objects[17]. Kalashnikov *et al.*'s QT-Opt trains a closed-loop, vision-based grasping policy using more than 580,000 real-world grasp attempts and achieves 96% success on previously unseen objects while learning behaviors such as regrasping and object repositioning[18]. Zeng *et al.* extend the action space itself by learning that pushing can support grasping, reporting 71.4% task completion compared with 42.9% for a grasping-only baseline in cluttered scenes[19]. Joshi, Kumra, and Sahin report improved grasping accuracy from a Double-DQN and Grasp-Q-Network when multi-view visual feedback is used instead of a single view[20].

What unites much of this literature is a shared assumption that is rarely stated outright: the gripper is treated as fixed, and most of the adaptive burden is placed on the policy[17-20]. This is a reasonable simplification for isolating algorithmic contributions, but it means the RL literature answers a different question from the morphology literature. Neither literature, on its own, explains what happens when a learned policy is paired with a compliant gripper rather than a rigid one. This is precisely the comparison Project Nazakat is designed to make.

Developmental and curriculum-based learning

A third literature asks whether the order of training experience matters independently of the algorithm used to exploit it. Curriculum-based deep RL has been proposed as a way to train adaptive robots through increasingly complex tasks, with reported benefits for convergence, training time, sample efficiency, and generalization[21]. Narvekar *et al.* provide a broader framework for curriculum learning in RL by classifying how tasks are generated, sequenced, and transferred between stages[22]. The foundational machine-learning work by Bengio *et al.* established the general principle of presenting training examples in a meaningful order, typically from easier to harder cases[23]. More recent manipulation work combines curriculum demonstrations with hierarchical RL so that high-level policies learn subtask sequences while low-level policies learn the individual skills[24], while reverse curriculum generation provides another way to expand training states from the goal outward[25].

The developmental-robotics literature provides the biological motivation for such staged learning. Wu *et al.* demonstrate a mobile manipulator that develops increasingly complex skills through simple and complex game modes[26]. Savastano and Nolfi present an incremental neurorobotic model of reaching and grasping in which developmental constraints help guide learning toward behaviors similar to those observed in infants[27]. Asada *et al.* argue that cognitive competence in embodied systems should be constructed progressively through interaction between body, brain, and environment[28]. Lungarella *et al.* similarly describe developmental robotics as a two-pronged field that uses robots both to study developmental processes and to apply developmental knowledge to robot design[29]. Oztop, Bradley, and Arbib model grasping as an extension of reaching through exploration, trial-and-error, and feedback[30], while Law *et al.* translate an infant developmental timeline into staged robot-shaping constraints that are gradually released as new abilities emerge[31].

Together, these studies show that curriculum learning and developmental robotics provide complementary pieces of the same idea. Curriculum-learning research supplies

mechanisms for ordering tasks and transferring knowledge, while developmental robotics supplies a biological rationale for the sequence itself. However, the reviewed literature provides limited evidence on whether an infant-inspired learning sequence improves manipulation robustness and generalization when tested under the same hardware and task conditions as a conventional RL curriculum.

Sim-to-real transfer

Because many of the systems above are trained substantially or entirely in simulation, the fidelity of that transfer is an important methodological issue. Scheikl *et al.* provide a concrete example in a deformable-manipulation setting: a visual RL policy trained in simulation was transferred to a real surgical robot through pixel-level domain adaptation without retraining and achieved 50% success on the real system using raw RGB images[32]. Zhao, Peña Queralta, and Westerlund organize the sim-to-real literature into domain randomization, domain adaptation, imitation learning, meta-learning, and knowledge distillation, and conclude that no single method works universally across differences in visual appearance, dynamics, sensing, and actuation[33]. Graule *et al.*'s SoMoGym addresses a related problem for soft robotics by providing standardized benchmark tasks and baseline RL results for continuum and soft robots[34]. Their results show strong performance from SAC across the benchmark suite, while other methods required more tuning, directly relevant to the present study's own choice of SAC as the underlying RL algorithm.

These studies highlight a direct methodological concern for any comparison involving rigid and compliant hardware. Simulation does not represent all forms of contact and deformation with equal fidelity, so the transfer behavior of a soft system may differ from that of a rigid system. For that reason, sim-to-real performance should be reported as part of the experimental result rather than treated only as a nuisance variable[32-34].

Synthesis and research gap

Read together, these four literatures leave a structured gap rather than a simple omission. The morphology literature shows that compliant and bioinspired grippers can improve adaptability and fragile-object handling, but it has not isolated those benefits from the controller paired with the hardware[5-14]. The RL-for-manipulation literature shows that learned policies can reach high grasp reliability, but much of this work uses fixed rigid hardware[15-20]. The developmental-learning literature provides a biologically grounded rationale for staged training and mechanisms for implementing curricula[21-31], yet it provides limited evidence on whether an infant-inspired sequence improves manipulation robustness and generalization under controlled experimental conditions. The sim-to-real literature further shows that transfer introduces additional differences between simulated and physical systems that should be measured explicitly[32-34].

We did not identify a study in this review that varies gripper morphology and learning curriculum independently within a single controlled design. Most prior work improves one factor while holding the other fixed, for example a soft gripper with a fixed heuristic controller, or a learned policy on a fixed rigid gripper. This leaves an important question unanswered: whether a robust grasp succeeds because the hand adapts physically, because the policy learns to compensate, or because the two properties interact in a way that neither can produce alone. Project Nazakat addresses this directly through a 2 × 2 factorial design, rigid versus soft (cephalopod-inspired) gripper crossed with standard versus developmental (infant-milestone-based) RL curricula, evaluated across fragile, deformable, irregular, and slippery objects, with the simulation additionally supporting evaluation under platform-disturbed conditions as a final curriculum stage. This is, to the best of the present review, a controlled attempt to separate the relative and combined contributions of compliant morphology and biomimetic learning to manipulation robustness on a single standardized benchmark.

Figure 1 summarizes this progression from the literature reviewed above through to the experimental design detailed in the following section.

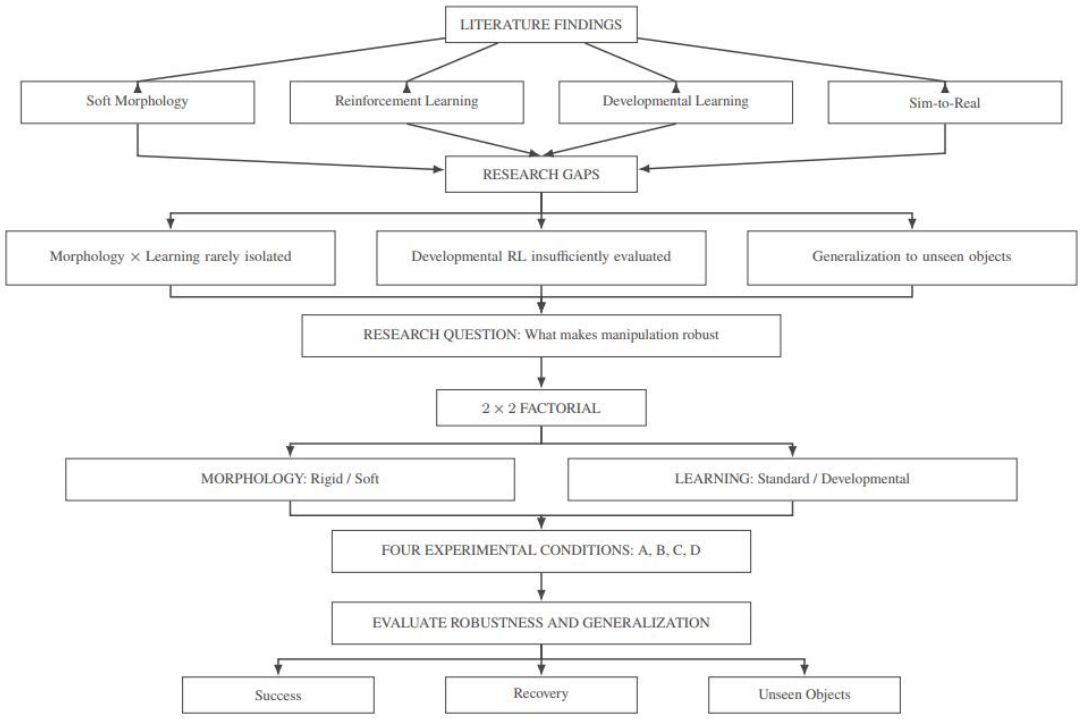


Figure 1. From literature findings to experimental design (Sections 2.6 and 3).

MATERIALS AND METHODS

Simulation environment

All experiments run in a custom physics simulation composed of three subsystems. A six-wheel-contact rover chassis model operates under Mars gravity (3.71 m/s^2), a value consistent with NASA's published planetary fact sheet[35], with per-wheel traction limits derived from terrain-dependent friction, slope-adjusted normal forces, a motor torque and power curve, aerodynamic drag, battery consumption, and rollover dynamics triggered by wheel lift-off. Terrain elevation is loaded from the real NASA Mars Orbiter Laser Altimeter (MOLA) dataset at 128 pixels per degree, used to generate physically grounded platform disturbances for the curriculum stage that trains manipulation under base motion. A 4-DOF serial-link arm (base yaw, shoulder pitch, elbow pitch, wrist pitch) is mounted on the chassis, using standard Denavit-Hartenberg forward kinematics, a numerical Jacobian, and damped least-squares inverse kinematics; joint motion is driven by a per-joint PD controller with a hard velocity limit (1.5 rad/s). Two

interchangeable gripper models share one abstract interface so morphology can be varied without touching the arm, RL environment, or training code:

Rigid: a two-point pinch model (following the standard parallel-jaw simplification of Mahler *et al.*[17]), 1 cm alignment tolerance, 40 N maximum grip force, 80 N/s force ramp rate. Objects are crushed if grip force exceeds the object's fragility limit before the required holding force is reached.

Soft: a 6-pad distributed-contact model inspired by cephalopod arm morphology. Each pad behaves as a damped spring with its own deformation limit; commanded force distributes across engaged pads, so peak contact force is much lower than the rigid gripper's two-point pinch at the same total force, and misalignment degrades gracefully rather than failing outright.

Experimental design

A controlled 2 × 2 factorial design isolates the contribution of gripper morphology and learning strategy, summarized in Table 1.

Table 1. The four experimental conditions.

	Standard RL	**Developmental RL**
Rigid Gripper	A. Baseline	C. RL contribution
Soft Gripper	B. Hardware contribution	D. Full system

As shown in Table 1, A-vs-B isolates the effect of gripper compliance alone; A-vs-C isolates the effect of the curriculum alone; A-vs-D measures the combined effect.

RL formulation

The observation space comprises a total of 29 dimensions, characterized by joint angles accounting for 4 dimensions and joint velocities accounting for another 4 dimensions. Other values include the world position of the end effector, the position of the target object in 3D space, the relative offset of both positions, the current grip force, and the state of the gripper status accounted for by 6 variables, and the rover state composed of 5 values. The action space accounts for 5 dimensions and is limited between -1 and 1. Four values in the action space relate to the angles of the joints while one of the values refers to the grip force.

The standard reward (Conditions A/B) is deliberately plain, with no shaping beyond a small time cost, so any generalization gap observed against C/D is attributable to the curriculum rather than to a richer reward:

$$r = -0.01 + \begin{cases} +1.0 & \text{gripper HOLDING} \\ -5.0 & \text{gripper CRUSHED} \\ -1.0 & \text{gripper DROPPED} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The developmental curriculum (Conditions C/D) trains six stages in sequence, each initialized from the previous stage's converged policy, mapped to human infant motor-development milestones [Table 2].

Table 2. Developmental curriculum stages and reward signals.

Stage	Milestone	Reward Signal
1	Exploration	+1 for visiting a novel 3 cm voxel of end-effector workspace (first visit only), -0.01 otherwise
2	Contact	+1 for contacting a novel angular sector of the object surface (12 sectors), +0.05 repeat contact, -0.01 out of range
3	Grip Discovery	2.0 × (displacement this step) +0.1 × (total displacement from spawn), -5.0 if crushed

Stage	Milestone	Reward Signal
4	Force Learning	+1.0 while holding, minus 0.05× (force in excess of the true minimum required); -5.0 crushed, -0.5 dropped
5	Dexterity	Same overshoot penalty as Stage 4, plus one-off +5.0 on first reaching HOLDING, +0.5/step held, -0.02 time cost
6	Environment Adaptation	Identical reward to Stage 5. Difficulty comes entirely from rover-motion disturbance during the grasp

CRUSHED ends an episode immediately; DROPPED and NO CONTACT are recoverable within the same episode.

Object Sets

The objects that were utilized during training are as follows: cube weighing 0.30 kg, ball weighing 0.20 kg, and cylinder weighing 0.25 kg are unbreakable in order to make sure that the manipulation skill is not being confused with fragility. Test objects: unseen cube and cylinder (the same shape group but different weight), rock and toy (irregular geometry, approximated through lower friction and a larger effective radius rather than true non-convexshape); egg surrogate (with fragility of 5 N) and paper cup (fragility level of 8 N); sponge and foam ball are a couple of deformable but unbreakable objects that still require a certain holding force during manipulation; smooth bottle has 0.15 value for friction coefficient.

Training Procedure

Soft Actor-Critic (SAC)[36] trains one policy per condition (and, for C/D, one policy per curriculum stage, each resumed from the previous stage's final weights), using Stable-Baselines3's default MlpPolicy architecture. All hyperparameters, identical across conditions, are listed in Table 3. SAC was chosen in part because Graule *et al.*'s SoMoGym benchmark suite found it to be the strongest-performing algorithm among those tested for soft-robot control tasks[34], directly relevant to Condition B/D's soft-gripper morphology.

An empirical correction applied uniformly across all four conditions: an initial run under the raw environment (one action per physics step) produced a flat reward for every training episode regardless of budget. Direct measurement showed that 400 episodes (200,000 environment steps) of uniform-random exploration never once brought the end effector within 5 cm of the target object. The cause was that the action space commands raw joint target angles rather than an IK-solved end-effector target, and the arm's PD controller is rate limited; at a 0.01 s physics step, a target re-sampled every step never gives the controller time to converge before being replaced. The fix, applied identically across all four conditions to preserve the fairness of the 2 × 2 comparison, is a standard control-decimation (action-repeat) technique: each sampled action is held for 50 consecutive physics steps before the policy is re-queried. This does not alter the reward function or the meaning of any action value, only the decision frequency. SAC's random-exploration warm-up was also raised from Stable-Baselines3's default of 100 timesteps to 20,000.

With an episode length of 500 physics steps and an action repeat of 50, each episode contains 10 policy decisions; the 3.0 s continuous-hold success criterion (Section 3.6) therefore spans 300 physics steps, or six consecutive decisions. Rewards and terminal events are computed at every physics step. The action-repeat wrapper sums the per-step rewards over the 50 repeated steps into the reward for a single policy decision, and ends the decision early if the episode terminates (CRUSHED) or times out partway through. Every per-step reward term, including the Stage 4/5 force-overshoot penalty and Stage 5's +0.5 hold reward, is therefore charged once per physics step, that is, up to 50 times per policy decision; this is what allows a single miscalibrated hold to accumulate episode rewards of several hundred below zero. The evaluator likewise steps the environment at physics-step resolution, re-querying the policy every 50 steps, and records the longest run of consecutive physics steps spent in HOLDING; an episode counts as a success if that run reaches 300 steps (3.0 s).

Table 3. Training hyperparameters (identical across conditions).

Parameter	Value
Algorithm	SAC
Policy network	SB3 default MlpPolicy
Action repeat	50 physics steps/decision
Learning starts	20,000
Episode length	500 physics steps (5.0 s sim time; 10 decisions)
Timestep (*dt*)	0.01 s
Seeds	0, 1, 2 (three independent runs per condition)

Evaluation protocol

Each trained policy is evaluated with exploration noise disabled (deterministic policy, reflecting deployed behavior) for 20 episodes per object across all 12 objects (240 episodes total per condition per seed). Eight metrics are computed: (1) grasp success rate, the percentage of episodes holding the object continuously for ≥ 3.0 s; (2) time to successful grasp; (3) RL convergence speed, the episodes for a rolling 100-episode window to reach ≥ 80% success on training objects, where a training episode counts as a success if the gripper is HOLDING at its final step (a looser criterion than the 3.0 s hold used in evaluation, logged during stochastic training); (4) failed grasp count per 20 attempts; (5) object drop rate; (6) average grip force; (7) generalization score, the pooled success rate across the 9 held-out test objects; (8) recovery rate, the percentage of drop events followed by a successful re-grasp within 1.5 s.

RESULTS

Total training budget per condition

The original plan gave C/D six stages times 100k timesteps, 600k total, deliberately matching A/B's 600k so total training compute was held constant across the 2 × 2 design. In practice, both curriculum conditions required substantially more: Condition C used 1,500,000 timesteps (Stages 1 through 4: 100k, 600k, 400k, 400k) to produce its reported checkpoint, and Condition D used 1,100,000 (Stages 1 through 3: 100k, 600k, 400k), 2.5 times and 1.83 times A/B's budget, respectively (Sections 4.4 and 4.5). Decision: A/B's training was not extended to match. Two considerations support this. First, the asymmetry already favors C and D, not A/B; if additional budget alone explained performance differences, C and D would be expected to outperform A and B, not underperform them as both did. Second, Condition A's own training curve (Section 4.2) showed diminishing returns well before 600k timesteps without reaching convergence, making it unlikely that additional budget would qualitatively change its result. Actual budgets are reported transparently per condition throughout Section 4 rather than assumed equal; the consequences of this mismatch for the factorial comparison are discussed in Section 6.

Condition A: Rigid gripper + standard RL (Baseline)

Trained to 600,000 decision-steps. A ten-window bucketing of the first 200,000-step segment of training (seed 0) showed a clear, monotonic improvement: successful episodes per 2,000-episode window rose from 3 to 5 in the first half of training to 10 to 18 in the second half, and mean episode reward rose from approximately -4.8 to -3.1 over the same span, confirming the policy learned throughout rather than stalling at the initial reward floor. Table 4 reports the pooled seed-0 evaluation.

Table 4. Condition A pooled results, seed 0.

	n	Success	Avg. Force	Crushed
Train objects	60	20.0%	35.77 N	0
Test objects (generalization)	180	11.67%	35.05 N	28

RL convergence speed was not reached (the rolling 100-episode success rate never crossed the 80% threshold within budget), expected for this condition, and, as Sections 4.4 and 4.5 show, Condition C never reached it either while Condition B did (Section 4.3), making convergence itself, not just final accuracy, part of what distinguishes the four conditions. Table 5 breaks the held-out test results down by object.

Table 5. Condition A per-object results, held-out test set, seed 0.

Object	Success	Crushed	Note
cube unseen	15.0%	0	N/A
cylinder unseen	20.0%	0	N/A
rock	0.0%	0	irregular geometry defeats alignment
toy	20.0%	0	N/A
egg surrogate	0.0%	14/20	70% crush rate
paper cup	0.0%	14/20	70% crush rate
sponge	25.0%	0	best-performing test object
foam ball	25.0%	0	best-performing test object
smooth bottle	0.0%	0	never sustains contact

Condition A generalizes moderately to rigid objects resembling its training set (15 to 20% on cube/cylinder variants and the irregular toy object) and performs best on the two deformable objects (25%), which tolerate imprecise alignment more forgivingly than a tightly-toleranced rigid grip. It fails completely on the irregular-geometry object (rock) and the slippery object (smooth-bottle). Most notably, average grip force is a near-constant 35 to 36 N (close to the 40 N hardware maximum) regardless of which object is handled. The policy learned a single, blunt squeeze-hard strategy that generalizes to unbreakable training objects but crushes fragile test objects on 70% of attempts, with all 28 crush events in the test set concentrated entirely in egg_surrogate and paper_cup. This is exactly the failure mode the project's design predicts for an unshaped, standard-RL, rigid-gripper baseline.

A second run with an independent random seed (seed 1) reached only 5.0% train success and 2.22% generalization, roughly 4 to 5 times lower than seed 0, with a correspondingly lower crush count (2/180 *vs.* 28/180), consistent with a less aggressive, less-developed version of the same blunt-force strategy rather than a qualitatively different failure mode.

A third run (seed 2) reached 0.0% on both train and generalization, the same clean floor Condition C reaches by a different mechanism. The stochastic training curve shows real, if noisy, improvement over the run (successes roughly doubling from the first half to the second), so this is not the flat, structurally-unlearnable pattern the original action-repeat bug produced; the deterministic policy simply never crystallized a version of the learned strategy reliably enough to reproduce it across 240 real evaluation episodes, with grip-force and drop-rate fields entirely blank because the gripper never once left NO_CONTACT deterministically. Unlike seed 1's still-climbing curve at the same 600,000-step mark, seed 2's success rate had already plateaued in its final quarter, so it was accepted as final rather than extended (see Section 6 for how this decision was applied across conditions). This gap is discussed alongside the other three conditions' own seed variance in Section 4.6.

Condition B: Soft gripper + standard RL

Trained to 600,000 decision-steps under identical hyperparameters to Condition A [Table 3]. The gripper is the only variable that differs between the two conditions, making this a clean, isolated morphology comparison. Unlike Condition A, seed 0 of this condition reached the paper's RL convergence-speed threshold: a rolling 100-episode success rate of ≥ 80% was reached at episode 15,717. Tables 6 and 7 report the pooled and per-object seed-0 results.

Table 6. Condition B pooled results, seed 0.

	n	Success	Avg. Force	Crushed
Train objects	60	100.0%	30.08 N	0
Test objects (generalization)	180	85.0%	29.83 N	20

Table 7. Condition B per-object results, held-out test set, seed 0.

Object	Success	Crushed	Note
cube unseen	95.0%	0	N/A
cylinder unseen	100.0%	0	N/A
rock	90.0%	0	N/A
toy	100.0%	0	N/A
egg surrogate	0.0%	20/20	100% crush rate
paper cup	100.0%	0	full success, 8 N fragility
sponge	100.0%	0	N/A
foam ball	100.0%	0	N/A
smooth bottle	80.0%	0	N/A

Compliant morphology alone, with no change to the learning strategy, produced a dramatic improvement over Condition A on this seed: 100% training success (*vs.* 20%) and 85% generalization (*vs.* 11.67%), and this is the first condition to reach the convergence threshold at all. Average grip force is also slightly lower (approximately 30 N *vs.* Condition A's 35 to 36 N), consistent with the softer gripper's larger effective tolerance allowing a somewhat gentler learned strategy. The crush pattern is the most specific finding: Condition B fully solves paper_cup (8 N fragility, the same object class Condition A struggled with) but still crushes egg_surrogate (5 N fragility, the single most fragile object in the benchmark) on 100% of attempts. This is a more precise claim than a general statement that soft grippers handle fragile objects better; the data indicates compliance alone raises the fragility ceiling the gripper can handle to somewhere between 5 N and 8 N, rather than eliminating crush risk universally.

Seed 1 of this condition is the single most important data point in this paper's multi-seed effort. Its stochastic training curve showed clear, monotonic improvement throughout its entire 600,000-step budget (non-floor episode rate climbing from 2.75% in the first 6,000-episode bucket to 39.1% in the last, still rising when the budget ended), yet its deterministic evaluation at that point returned only 10.0%/7.22%, a tenfold gap from seed 0. This matched a pattern seen earlier with Condition A: the stochastic rollout, exploration noise included, can run well ahead of the deterministic policy early in training. Extending seed 1 by a further 400,000 steps (1,000,000 total, 67% more than seed 0 needed to converge) raised the deterministic result to 20.0%/13.89%, real progress, but the extended run's own stochastic success rate plateaued near 55 to 56% in its final quarter rather than continuing to climb, the signal used throughout this paper to distinguish true stalls from checkpoints read too early. Seed 1 was therefore accepted at 20.0%/13.89% as its final result rather than extended further.

A third run (seed 2) reached 15.0%/10.56% at the standard 600,000-step budget. Its stochastic success rate (32.0% in the final in-training window) had not obviously plateaued, so, applying the same still-climbing criterion used for seed 1, it was extended by a further 400,000 steps. Unlike seed 1, this extension made the result worse rather than better: the deterministic evaluation fell to 10.0%/6.11%, and the in-training stochastic success rate itself declined slightly over the extension (32.0% to 29.0%), confirming this was a real regression during continued training rather than deterministic-eval noise. Exactly as with Condition C's Stage 4 learning-starts regression, a periodic checkpoint from before the regression, saved automatically at the 600,000-step mark, was still available; re-evaluating it directly reproduced the pre-extension result almost exactly (15.0%/10.56%, avg. grip force and crush counts matching within eval-order noise), confirming it was not a fluke of the original run. Seed 2's final result is therefore 15.0%/10.56%, the recovered pre-extension checkpoint, following the same retain-the-better-known-state principle already applied to Condition C. The extension attempt itself is reported here as a second, independent instance of

continued training regressing a good checkpoint under this project's reward formulations, not as a result in its own right: the still-climbing stopping rule that correctly identified seed 1 as under-trained did not reliably predict seed 2's outcome, and the two seeds of the same condition were, as a direct consequence, extended under an identical rule with opposite results. This asymmetry, and what it implies about trusting a single stochastic-curve heuristic to decide when a run is truly finished, is discussed further in Section 6.

Condition C: Rigid gripper + developmental RL

Going through a notable training history of this condition is worthwhile not only because one needs to summarize it but mainly because it unfolded in such a way that makes it one of the noteworthy results of this research. All checkpoint values provided below were directly measured and not obtained as a result of any inference.

Stage 1 was successful as far as training is concerned. At that initially theoretically determined landmark of 100,000. This was to be expected as at this stage, the reward mechanism known as voxel-novelty does not take into consideration the spatial position of the object, thus gaining general motor skills and not teaching the policy how to aim targets.

Stage 2 is where the situation became complicated. The same originally planned number of steps (100,000) was used for stage 2 and the attempt of the policy to learn its objective failed as measuring proves that the mean value of closest approach was 19.28 cm which is statistically identical to the outcome of pure random exploration (Section 3.5). Extending training to 600,000 steps total actually fixed this, and Table 8 shows the trajectory: mean distance dropped to 13.57 cm by the end, with contact rate climbing from 1/100 to 13/100 along the way.

Table 8. Condition C, Stage 2 extended-budget diagnostic, seed 0.

Checkpoint	Mean	Median	Best Ep.	Contact Rate
100k (original)	19.28 cm	18.45 cm	4.32 cm	1/100
450k	16.38 cm	16.14 cm	1.05 cm	11/100
520k	16.70 cm	15.55 cm	3.40 cm	12/100
600k (final)	13.57 cm	12.39 cm	0.0000 cm	13/100

One checkpoint 70,000 steps into that extension (450k to 520k) looked flat and got flagged at the time as a possible plateau. It wasn't. The entirety of the current practice illustrates that the flat sections merely provided some noise on top of actual, progressive improvement, serving as a good reminder of the fact that if a checkpoint seems stuck in the future, it might actually not be.

Stage 3 confirmed the results from Stage 2 and with a proactive budget of 400,000 steps, showed a truly smooth and improving tendency. Initial evidence shows that experiences that achieved the HOLDING status at least once, which is significantly less challenging than the threshold of the requirement for the sustained-hold, increased from 8 to 11 per bucket of 4,000 episodes at the start of the training to 23-25 per bucket at the finish.

Stage 4 is where the curriculum's central problem first showed up. Warm-started from Stage 3 and given its own 400,000-step budget, it reached a mean closest-approach of 11.52 cm and a 17/100 contact rate, real progress carried over from the previous stage. Trying to extend that run further, though, surfaced a real bug in how resuming was set up: continuing the same stage past its planned budget was still using the large learning_starts value (20,000) meant for a genuine stage transition, which threw away 20,000 steps of the already-loaded policy's judgment in favor of pure random actions against an empty replay buffer. The result was a measurably worse policy, mean closest-approach regressed to 20.06 cm, contact rate cut by more than half to 5/100. The lesson is simple in hindsight: learning_starts should stay small when a stage is just being extended past its own budget, and only be large when the reward function itself is

genuinely changing between stages. We corrected the bug and re-ran it, and while gradient updates restarted properly from step one, the run did not recover any further improvement within its own budget, so we kept the original, known-good 11.52 cm, 17-of-100-contact checkpoint as Stage 4's official result rather than the regressed one.

Stage 5 is one at which things did not just stagnate but actually regressed. With the same budget of 400,000 steps, the system was found to have declined in all parameters [Table 9]: the mean closest approach worsened to 16.30 cm, the contact rate dropped to 8/100, and the success of the reaching_holding was still at 0/100 in any training run.

Table 9. Condition C, Stage 4 (retained) *vs.* Stage 5 (final), seed 0.

	Stage 4	Stage 5	Change
Mean closest-approach	11.52 cm	16.30 cm	worse
Median	11.57 cm	16.86 cm	worse
Contact rate (≤ 6 cm)	17/100	8/100	worse
reached holding (100 eps)	0/100	0/100	N/A

The reason for this is explained by the training curve. The average success rate for the Stage 5 per-episode was declining over the run. The success rate started at 17 in 4,000-episode bucket (inherited from Stage 4), and dropped to between 3 and 8 for the rest of the run, finishing at 4. The rewards per episode dropped to -530 in some episodes while remaining below the 'do-nothing' floor of -10. The reason is pretty simple once you see it: Stage 5 has the same uncapped force-overshoot penalty of Stage 4 (-0.05 times the excess grip force every time the gripper holds), which adds up over the course of one miscalibrated grip, while the 'do-nothing' does not have such penalty. It is a genuine, exploitable flaw in how the reward is shaped: the policy learned to avoid gripping altogether, and the steadily declining per-bucket success count over a fully completed training run is direct evidence of exactly that.

Stage 6 was excluded from our experimentation as it incorporates the same reward function as Stage 5 unchanged, adding only a motion disturbance to the environment, thus we did not expect it to regain the output of the previous stage. Stage 4 was designated the best-performing situation in the context of the training sequence, thus its outcome was acknowledged as the official outcome of Condition C.

That final result is about as clean a failure as you can measure: 0.0% across every one of the 12 objects individually [Table 10], with not one episode out of the full 240-episode evaluation ever reaching HOLDING, matching the 0-of-100 diagnostic exactly.

Table 10. Condition C final evaluation, Stage 4 checkpoint, seed 0.

	n	Success	Crushed
Train objects	60	0.0%	0
Test objects (generalization)	180	0.0%	0

RL convergence speed does not apply here either, the threshold was never reached, same as Condition A.

Seed 1 landed at the identical 0.0%/0.0% result, but by a somewhat different, and in one respect more revealing, route. Its Stage 2 actually outperformed seed 0 at the same 600,000-step budget (mean 7.07 cm *vs.* 13.57 cm, contact rate 42/100 *vs.* 13/100) despite starting slower, and its Stage 3 racked up more total training-episode successes than seed 0 did at the same stage (1.07% *vs.* 0.44% of all episodes). But Stage 4's final evaluation told a different story: average grip force of 6.05 N on train objects and 2.89 N on test objects, with object drop rates of 100% and 66.67% respectively, figures that were entirely blank for seed 0, since that gripper never once left NO_CONTACT. Seed 1's gripper, in other words, did occasionally reach HOLDING, matching its own

2-of-100 diagnostic reading, and then dropped the object every single time it did. That is the first direct behavioral evidence of the grip-then-release mechanism the Discussion (Section 5) argues for on theoretical grounds. Seed 0's total absence of any holding event could only support that theory indirectly. Seed 1 shows the actual grip-and-immediately-drop pattern the reward's shape predicts, happening in practice.

All of these results hold significance. It becomes very clear from Condition C that stages 1 to 3 of the curriculum were successful in that they imparted measurable competencies to the policy, well beyond what Condition A's flat 600,000-step run ever achieved. Stages 4 and 5 exhibit a clear problem, as they use excessive penalties for invalid steps that may lead to the learning of incorrect gripping skills, since the policy did not learn the correct force in the first place. Thus, we are able to formulate a very precise statement about which parts of the entire program are flawed. This finding, confirmed first on seed 0, now also applies to seed 1.

A third seed, seed 2, also landed at 0.0% on both metrics, but got there through a genuinely different mechanism than either of the first two. Its Stage 2 was initially trained at the wrong budget entirely, a command error left it at 100,000 steps instead of the required 600,000, producing a checkpoint statistically indistinguishable from the very first under-budgeted Stage 2 attempt described earlier in this section. Once we caught that and retrained it properly to the full 600,000 steps, this seed's Stage 2 still converged to a weaker state than either of the other two, a 7-of-100 contact rate, against 13/100 for seed 0 and 42/100 for seed 1. Stage 3, warm-started from that weaker foundation, showed a real but short-lived improvement early on, training-episode successes climbed to 10 per 4,000-episode bucket somewhere between 40,000 and 120,000 steps into the stage, before falling back toward zero for the rest of the run. Neither the earliest checkpoint we had saved nor the final one showed any deterministic holding at all, 0 out of 100 on both, which tells us the decline had already fully set in before any usable intermediate state got preserved. Given a foundation with no

measured grip-discovery capability whatsoever, there was no reasonable case for proceeding to Stage 4, so we stopped the curriculum at Stage 3 for this seed. The resulting evaluation matched seed 0's floor exactly, 0.0% on both train and generalization, but for a genuinely different reason: seed 0's collapse came from the Stage 4/5 overshoot penalty tearing down a Stage 3 policy that had actually been working. This seed never built a working Stage 3 policy to tear down in the first place. We report this as its own distinct finding rather than folding it into seed 0's result, since it shows something seed 0 alone could not: Condition C's fragility is not limited to the one Stage 4/5 reward flaw already identified. A sufficiently weak early-stage foundation can produce the exact same floor on its own, without the policy ever reaching the stage where that specific flaw even comes into play.

Because Stage 2's difficulty only depends on the distance between the end effector and the object, not on which gripper is attached, we applied the same extended budget to Condition D's Stage 2 pre-emptively, rather than waiting to rediscover the same failure there independently.

Condition D: Soft gripper + developmental RL

Stage 1 (Exploration) trained cleanly on the standard 100,000-step budget, mean reward climbing from approximately 190 to 308, matching Condition C's own Stage 1 (this stage's reward is gripper-independent).

Stage 2 (Contact), given the standard 600,000-step budget pre-emptively, produced a dramatic, clean result, markedly better than Condition C's equivalent stage on every metric [Table 11].

Table 11. Condition C *vs.* D, Stage 2 (final checkpoints), seed 0.

	C (Rigid)	D (Soft)
Mean closest-approach	13.57 cm	3.93 cm

Median	12.39 cm	3.97 cm
Contact rate	13/100	68/100

The training curve was the cleanest seen across the project, showing monotonic improvement, no false-plateau scare, no diagnostic rescue needed: non-floor episodes rose from 144/6,000 early to 4,231/6,000 (36.4%) by the end, mean reward crossing from -4.89 to +3.53. This is direct evidence that gripper morphology affects not just final performance (already shown by Condition B) but trainability itself.

Stage 3 (Grip Discovery), at the standard 400,000-step budget, produced the strongest result in this project's entire curriculum history (Table 12).

Table 12. Condition D, Stage 3 final checkpoint and banked evaluation, seed 0.

Metric	Value
reached holding (100 eps, deterministic)	65/100
sustained success (≥ 3 s, 100 eps)	32/100
Training-episode HOLDING rate (40k eps)	52.38% (*vs.* C's 0.44%)
Eval: train success (60 eps)	35.0%
Eval: generalization (180 eps)	30.0%

This checkpoint was evaluated in full and banked specifically as insurance before proceeding to Stage 4, given what happened to Condition C at this exact point in its own curriculum. It already exceeds Condition A's final, fully-trained seed-0 result (20.0%/11.67%) at only the third of six stages.

Stage 4 (Force Learning), at the standard 400,000-step budget with intermediate checkpoints actively monitored, produced a partial version of Condition C's failure mode [Table 13].

Table 13. Condition D, Stage 3 (banked) *vs.* Stage 4 (final), seed 0.

	Stage 3	Stage 4	
Contact rate	93/100	83/100	mild decline
reached holding	65/100	45/100	meaningful decline
sustained success	32/100	0/100	collapse

sustained success is the quantity the evaluation protocol's grasp-success criterion (Section 3.6) is directly built on, and it collapsed to zero despite reached-holding remaining substantial (45/100), the same mechanism identified in Condition C: the uncapped overshoot penalty accrues the longer a hold continues, incentivizing a brief grip-and-release over commitment through to a 3-second sustained hold. This is a second, independent confirmation of the same reward-design flaw on different hardware, strengthening the claim that it is a general property of the reward formulation rather than a gripper-specific artifact.

The curriculum was stopped at Stage 3. Stages 4 through 6 all carry the same unrepaired reward-overshoot formulation that, in every continuation tested here, degraded an already-performing checkpoint. Condition D's final result is Stage 3's checkpoint: 35.0% train and 30.0% generalization.

This is an important finding for the paper's central research question. On seed 0, Condition B (soft + standard RL, no curriculum) reaches 100%/85%, substantially outperforming Condition D (soft + developmental) at 35%/30%. For this curriculum design, once the Stage 4/5 flaw is accounted for, the developmental curriculum subtracts value rather than adding it on top of compliant morphology, the opposite of what a

clean multiplicative story would predict (Section 4.6 examines how consistently this holds across seeds).

Seed 1 of Condition D, stopped at the same Stage-3 point for the same documented reason, reached 15.0%/10.0%, well under half of seed 0's Stage-3 result on both metrics (reached-holding 34/100 *vs.* 65/100; sustained-success 12/100 *vs.* 32/100), a smaller relative gap than Condition A or Condition B showed between their own seeds at the time. This is discussed further in Section 4.6.

Seed 2, given the standard budget at every stage after an initial command error required Stage 2 to be retrained at its correct 600,000-step budget (the same class of mistake documented for Condition C in Section 4.4, caught and corrected the same way), reached a weaker Stage 2 foundation than either prior seed (contact rate 20/100, versus 68/100 for seed 0 and a comparable figure for seed 1). Despite this, Stage 3 recovered well: reached-holding of 34/100 and sustained-success of 14/100, essentially matching seed 1 despite starting from a substantially weaker Stage 2 state. The final evaluation, 15.0% train and 14.44% generalization, lands close to seed 1's result. Unlike Condition C, where a similarly weak Stage 2 foundation for seed 2 directly capped what Stage 3 could achieve, Condition D's Stage 3 appears more able to compensate for a weak starting point given its full training budget, a difference worth noting though not overstating from a single instance.

Multi-seed summary

Table 14 reports mean ± sample standard deviation across three independent seeds for all four conditions, alongside each condition's individual seed values.

Table 14. All four conditions, three seeds, mean ± std.

Condition	Seed 0	Seed 1	Seed 2	Mean ± Std
A, train	20.0%	5.0%	0.0%	8.3 ± 10.4%

A, gen.	11.67%	2.22%	0.0%	4.6 ± 6.2%
B, train	100.0%	20.0%	15.0%	45.0 ± 47.7%
B, gen.	85.0%	13.89%	10.56%	36.5 ± 42.1%
C, train	0.0%	0.0%	0.0%	0.0 ± 0.0%
C, gen.	0.0%	0.0%	0.0%	0.0 ± 0.0%
D, train	35.0%	15.0%	15.0%	21.7 ± 11.5%
D, gen.	30.0%	10.0%	14.44%	18.1 ± 10.5%

Four observations follow, the fourth new to the three-seed picture and not visible with two. First, every condition except C shows substantial seed-to-seed variance, and this variance did not shrink with a third data point. Condition A's third seed reached an actual floor of 0.0% on both metrics, meaning its range now spans from a complete failure to 20.0% train success, too wide to summarize as a single multiplier. Condition B's spread widened rather than narrowed, from a 5 to 6 times swing across two seeds to 6.7 to 8.0 times across three (15.0% to 100.0% train, 10.56% to 85.0% generalization). Condition D remained the tightest of the three variable conditions, 2.3 to 3.0 times across its three seeds. Second, the direction of the morphology effect is unchanged across all three seeds when seeds are compared one-to-one: each B seed exceeds the A seed with the same index on generalization (85.0% *vs.* 11.67%, 13.89% *vs.* 2.22%, 10.56% *vs.* 0.0%), and each D seed likewise exceeds the corresponding C and A seeds. This does not hold for every cross-seed pairing: B's seed 2 (10.56%) falls just below A's seed 0 (11.67%), and D's weakest seed (seed 1, 10.0% generalization) also falls below A's seed 0. C's three seeds land at exactly 0.0%/0.0%, the only condition with zero variance, now confirmed a third independent time. Third, Condition B, the paper's strongest condition on seed 0, remains its least reproducible across all three seeds, a fact a single-seed or even a two-seed report would have understated rather than resolved, since the spread widened, not narrowed, with the third data point. Fourth, and only visible with three seeds: Condition C's exact zero across all three independent seeds is now a stronger, more specific result than it was with two. Two identical zeros could

plausibly reflect two seeds discovering the same degenerate optimum by a coincidence of similar initialization; a third independent zero, arrived at via a distinct route (a budget-command error compounding into an already-weak Stage 2 foundation, documented in Section 4.4, rather than the Stage 4/5 overshoot mechanism seen in the other two seeds), makes it far less plausible that Condition C's collapse is an artifact of any single seed or any single failure mechanism. Condition D's comparatively tighter seed spread is itself a secondary finding, unchanged in direction by the third seed: the curriculum's early stages, whatever their later-stage flaw, may produce more consistent intermediate competence across random initialization than standard RL does on the same soft-gripper hardware, even though their peak performance remains well below B's. Figure 2 plots every seed individually against its condition mean.

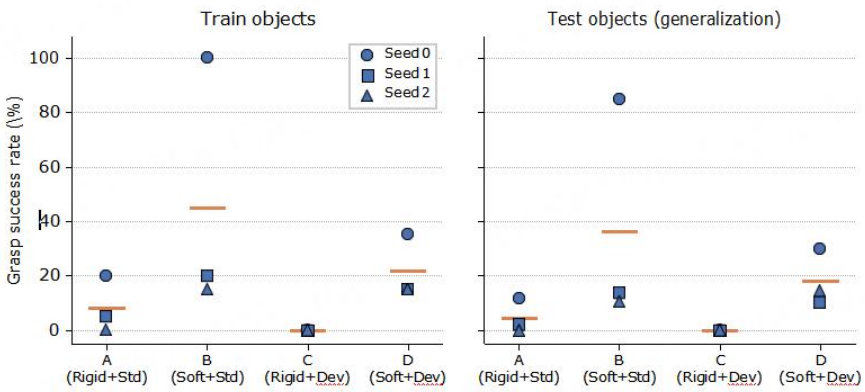


Figure 2. Grasp success rate by condition, all three seeds individually (markers: circle, square, triangle for seeds 0, 1, 2) against the condition mean (orange line). Condition B's spread is driven by one high outlier; Condition C's markers coincide at zero.

Figure 2 shows this same data visually. Condition B's spread is driven by a single high outlier, seed 0's 100.0%/85.0%, against two considerably lower seeds, rather than three seeds spread evenly across its range; Condition A's and D's spreads are, by contrast, more evenly distributed across their respective ranges.

Because the three seeds of each condition share seed indices, the factor effects can also be read seed by seed [Table 15]. This pairing is nominal, since different conditions consume random numbers differently, but it gives the distribution of each effect rather than only a difference of means. The morphology effect (soft minus rigid) is positive in all six seed-matched comparisons, under both learning strategies. The curriculum effect (developmental minus standard) is zero or negative on all three rigid-gripper seeds, but mixed on the soft gripper: strongly negative on seed 0, slightly negative on seed 1, and slightly positive on seed 2 for generalization. With only three seeds, conventional confidence intervals are not informative (a two-sided 95% t-interval on the mean A-to-B generalization difference spans roughly -57 to +121 percentage points), so we treat these results as directional evidence about the sign and consistency of each effect, not as precise effect-size estimates, and we base our conclusions on the three-seed means and seed-matched signs rather than on the best single run.

Table 15. Seed-matched factor effects, in percentage points (second condition minus first).

Effect	Metric	Seed 0	Seed 1	Seed 2	Mean
A→B (morphology, standard RL)	train	+80.00	+15.00	+15.00	+36.67
	gen.	+73.33	+11.67	+10.56	+31.85
C→D (morphology, developmental)	train	+35.00	+15.00	+15.00	+21.67
	gen.	+30.00	+10.00	+14.44	+18.15
A→C (curriculum, rigid)	train	-20.00	-5.00	0.00	-8.33
	gen.	-11.67	-2.22	0.00	-4.63
B→D (curriculum, soft)	train	-65.00	-5.00	0.00	-23.33
	gen.	-55.00	-3.89	+3.88	-18.34

Cross-condition comparison

Table 16 and Figure 3 compare the final seed-0 results of all four conditions.

Table 16. All four conditions, seed-0 final results.

Condition	Train	Gen.	Notes
A, Rigid+Standard	20.0%	11.67%	never converged
B, Soft+Standard	100.0%	85.0%	best overall
C, Rigid+Developmental	0.0%	0.0%	Stage 4/5 collapse
D, Soft+Developmental	35.0%	30.0%	stopped at Stage 3

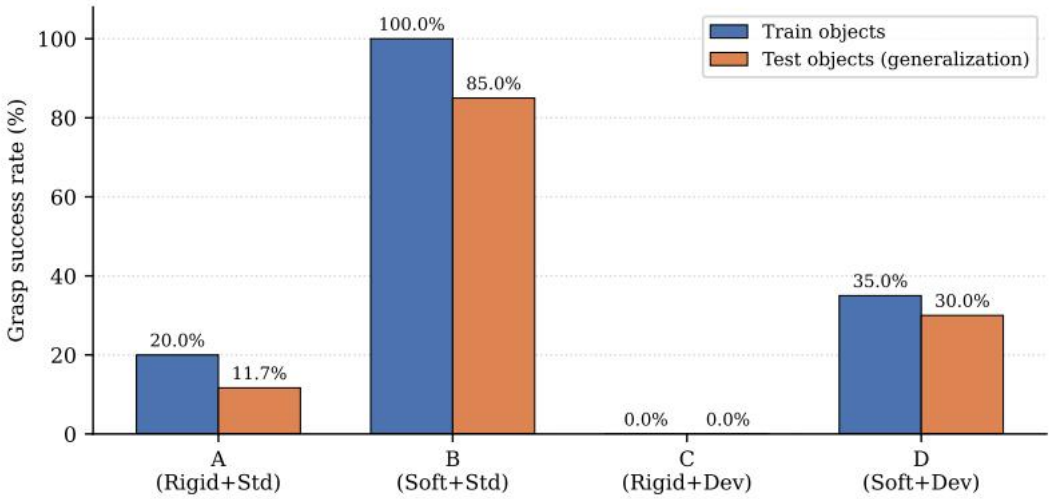


Figure 3. Grasp success rate by condition, train *vs.* generalization, seed 0. Condition B dominates; both curriculum conditions (C, D) underperform their non-curriculum counterpart on identical hardware.

A-vs-B (morphology alone) shows an enormous effect on seed 0: 11.67% to 85.0%. The seed-matched effect is positive on all three seeds but much smaller on seeds 1 and 2 [Table 15]; on the three-seed means, generalization rises from 4.6% to 36.5%. Compliant morphology produces the largest and most consistent improvement observed in this study.

A-vs-C (curriculum alone) shows that curriculum alone, as implemented, performs worse than or equal to the unshaped baseline on every seed once its later stages are included (0% *vs.* 11.67% on seed 0), despite demonstrably teaching real intermediate

skills (Stages 1 through 3) Condition A never developed. The final generalization score alone hides this; the intermediate diagnostics are necessary to tell the full story.

B-vs-D (curriculum's effect on top of morphology) shows curriculum makes things worse on top of morphology on average too (85.0% to 30.0% on seed 0; 36.5% to 18.1% on the three-seed means), for the same reason, although on seed 2 D slightly exceeds B.

The headline finding is that across every comparison, the confound is the same: the Stage 4/5 reward design, an uncapped per-tick force-overshoot penalty, is the dominant factor separating every curriculum condition's final number from what its own intermediate progress would predict, independently confirmed on two different grippers and, within Condition C, on two independent seeds, ruling out both a gripper-specific and a seed-specific explanation. Morphology helps consistently in this simulator. This curriculum's later stages hurt, and do so for a specific, identified, and fixable reason.

DISCUSSION

This study set out to determine whether robust robotic manipulation is driven more by compliant physical morphology, biologically inspired learning, or their combination, and whether these two factors add or multiply when combined. The results give a clear, if unexpected, answer for this simulator: compliant morphology produced the largest and most consistent effect. The soft gripper improved on the rigid gripper in every seed-matched comparison under both learning strategies [Table 15], and Condition B (soft gripper, standard RL) reached the best single result observed (85.0% generalization, seed 0) using less total training compute than either curriculum-trained condition, although its three-seed mean (36.5 ± 42.1%) shows that this peak is not typical. The developmental curriculum did not add to this on average: Condition D's mean generalization (18.1%) falls well short of Condition B's (36.5%) despite identical hardware and nearly double the training budget, and Condition C never exceeds

Condition A on any seed. Rather than adding to or multiplying with morphology's contribution, this curriculum, as implemented, subtracted from it on average.

This is not, however, evidence that staged learning is inherently unproductive. Both curriculum conditions' early stages (1 through 3) produced measurable competence their non-curriculum counterparts never developed under an equivalent unshaped reward: Condition C's Stage 3 training run reached a training-episode grasp rate roughly two orders of magnitude higher than Condition A ever approached, and Condition D's Stage 3 checkpoint, evaluated in isolation, reached 35.0%/30.0%, already ahead of Condition A's final, fully-trained seed-0 result. The failure is specifically localized: the moment Stage 4 introduces its force-calibration objective, an uncapped per-tick penalty proportional to excess grip force, both conditions' sustained-success rate collapses toward zero, independently and reproducibly, on two different grippers and, within Condition C, on two independent random seeds. A policy that has not yet learned to calibrate force correctly faces a reward landscape in which never attempting a grasp is a safer expected outcome than attempting one, since a single sustained hold under poorly-calibrated force can accumulate a penalty (observed as low as -530 in a single episode) that dwarfs any reward available for succeeding. Condition C's second seed makes this mechanism directly observable rather than merely inferred: its gripper did occasionally reach HOLDING, then dropped the object on every single one of those occasions (100%/66.67% drop rates, train/test), the grip-and-immediately-release pattern the reward's shape predicts, not just its absence. This is a specific, mechanistically identified property of one reward function's shape, not a general indictment of developmental curricula, but it is plausible, though outside what this study can directly confirm, that the same failure mode would recur in any staged-reward RL system that introduces an unbounded penalty term before the skill it penalizes deviation from has been reliably acquired.

Two directions suggested themselves for revising this specific reward. One, tested directly, was to reduce the overshoot penalty's coefficient fivefold (0.05 to 0.01), bounding the worst-case cost of an attempted grasp without removing the penalty's incentive structure entirely [Table 17]. Resumed from the same Stage 3 checkpoint and given the same 400,000-step budget as the original Stage 4 run, the softer coefficient produced a clearly different training signature: no episode reward fell below zero at any point in the run (versus a minimum of -530 under the original coefficient), and the per-bucket success rate stayed flat rather than declining, indicating the active avoidance behavior seen in Sections 4.4 and 4.5 was substantially reduced. However, the deterministic evaluation was unchanged: reached-holding remained exactly 0/100 and the full 240-episode eval returned 0.0%/0.0%, identical to the original coefficient's result. This is an informative negative result in its own right, not a dead end. It reveals that the avoid-gripping mechanism, while real (Sections 4.4 and 4.5), is not the sole bottleneck. Even C's original Stage 4 checkpoint, under the unmodified 0.05 coefficient, never once reached HOLDING deterministically either [Table 9], meaning some episodes' cumulative reward came from brief, noise-assisted touches during stochastic training that the noise-free deterministic policy could not yet reproduce, a distinct problem from active avoidance. Softening the penalty measurably fixed the avoidance dynamic without touching this separate precision bottleneck: reliably landing within the gripper's 1 cm alignment tolerance and commanding well-calibrated force at that exact instant, simultaneously, via a raw joint-angle action interface. This points more directly at the action-space limitation already noted in Section 6 than at reward-coefficient tuning as the more likely path to a full fix.

Table 17. Overshoot coefficient reduced fivefold (0.05 to 0.01), Condition C Stage 4, same starting checkpoint and budget.

	0.05 (Original)	0.01 (Tested)
Min training-episode reward	-530	0.00
Success-rate trend	declining	flat

Mean closest-approach	11.52 cm	10.29 cm
Contact rate	17/100	25/100
reached holding (100 eps)	0/100	0/100
Eval: train / generalization	0.0% / 0.0%	0.0% / 0.0%

A second direction, ramping the penalty's coefficient up gradually over the course of Stage 4 rather than applying either value at full strength from the first episode, was tested directly on both grippers (Condition C and Condition D, seed 0, both resumed from their respective Stage 3 checkpoints and given the same 400,000-step budget, coefficient rising linearly from 0.0 to the original 0.05 across the full budget). Checking only each run's final checkpoint would have understated what actually happened, particularly for Condition D, and the full picture required checking intermediate checkpoints as well, the same discipline applied throughout this paper [Table 18].

Table 18. Ramped overshoot coefficient (0.0 to 0.05 across Stage 4), both grippers, seed 0. Ckpt. columns are intermediate checkpoints at roughly a quarter and half of the stage's budget.

	C, Flat	C, Ramp	C, Ckpt.	D, Ramp	D, Ckpt.
Contact rate	17/100	17/100	0/100	61/100	74/100
reached holding	0/100	0/100	0/100	27/100	39/100
sustained success	0/100	0/100	0/100	0/100	19/100

On Condition C, the ramp changed nothing measurable at its final checkpoint, and an earlier intermediate checkpoint (roughly a quarter of the way through Stage 4's budget) showed no rescue either, remaining at 0/100 on every metric checked. On Condition D, the final checkpoint told an incomplete and actively misleading story on its own: it showed the ramp making the result worse than the flat coefficient (27/100 reached-holding *vs.* 45/100). An intermediate checkpoint, saved roughly halfway through Stage 4's budget, told a different and more interesting story: 39/100

reached-holding and, for the first time anywhere in this paper's Stage 4 or Stage 5 results under any reward variant tested, an actually nonzero sustained success, 19/100. The full 240-episode evaluation of this checkpoint confirmed this was real: 10.0% train success and 8.89% generalization, with 11 crush events, the first time crushing has appeared at all in this condition's Stage 4 attempts.

The training curve explains both the rise and the fall. With the ramp coefficient near zero early in training, the policy achieved real, high-reward holding success inherited directly from Stage 3's warm start, visible as a strongly positive mean reward in the first fifth of the run. As the coefficient climbed past the point where the policy's force calibration remained reliable, the same catastrophic-penalty dynamic identified in Sections 4.4 and 4.5 reappeared, episode rewards falling as low as -348 in single episodes, and continued training past the rescued checkpoint pushed the policy toward the same avoid-the-risk equilibrium documented throughout this paper, discarding the real capability the intermediate checkpoint had reached. This reframes the finding considerably from a clean negative result to an early-stopping problem: the ramp can produce real, holding-capable intermediate states, at least on the soft gripper, but continuing training past that point degrades rather than improves them, and nothing in the training signal available at run time (reward, success rate) flagged the rescued checkpoint as the best state to keep, the same blind spot already discussed for Condition C's original Stage 4 checkpoint and Condition B's seed 2 extension.

This finding should be read narrowly, not as a practical fix. The rescued Condition D checkpoint's 8.89% generalization remains well below Condition D's actual reported seed-0 result of 30.0%, reached simply by stopping the curriculum at Stage 3 and never attempting Stage 4 at all (Section 4.5). The ramp, even at its best discovered checkpoint, does not outperform not running Stage 4; it demonstrates only that some real force-calibration capability can survive Stage 4's reward under the right stopping point, on at least one gripper, not that doing so is currently worthwhile. Condition C showed

no equivalent rescue at the one intermediate checkpoint checked, so whether this effect is specific to the soft gripper's more forgiving contact model or simply a matter of checkpoint timing that a different save interval would also have caught for Condition C remains an open question. Given this result on both grippers, softening or delaying the same penalty in different ways produces, at best, a fragile and hard-to-locate intermediate state rather than a reliable fix, strengthening the case made above that the underlying limitation sits in the action-space interface rather than in this specific reward term's shape or schedule. Any further revision to C and D's reward would still need to be evaluated with the same controlled rigor applied here, since changing only C and D's reward without a matching adjustment to how A and B's baseline already excludes force shaping entirely would reopen exactly the confound this study's design was built to avoid.

A secondary contribution of this study is methodological rather than empirical. The mechanism behind Conditions C and D's underperformance was identifiable at all only because training was monitored through intermediate, stage-specific diagnostics, direct measurement of end-effector precision and holding rate at multiple checkpoints, rather than through final aggregate success scores alone. A single end-of-training evaluation would have reported only that the developmental curriculum failed, with no way to distinguish a true failure of the staged-learning approach from a narrow, fixable defect in one stage's reward shaping. The distinction matters: the former would argue against developmental curricula as a strategy; the latter argues only against one specific implementation choice within one stage of one curriculum. Future work applying staged reward curricula to sparse-reward manipulation tasks would benefit from the same discipline, tracking task-relevant proxy metrics at the checkpoint level throughout training, not only at its conclusion.

The multi-seed results (Section 4.6) add a second methodological lesson in the same spirit. Condition B's seed 1 would have been reported as an outright failure

(10.0%/7.22%) had it been evaluated immediately after its planned 600,000-step budget without checking whether its stochastic training curve was still climbing. It was, and the deterministic policy caught up substantially with a further 400,000 steps. Reporting only the final deterministic number, without the intermediate stochastic trend that distinguishes a truly stalled run from a still-converging one, would have overstated Condition B's fragility to random seed just as easily as it could understate a stalled run's true state.

Returning to the research gap identified in Section 2.6, we did not find prior work reviewed there that varies gripper compliance and learning curriculum independently within a single controlled design, leaving open whether a successful grasp, when observed, reflects the hand adapting physically, the policy compensating computationally, or real synergy between the two. This study's controlled comparison addresses that question for this simulator, if not in the direction the developmental-robotics literature's framing might suggest: for the manipulation task and reward formulations examined here, physical compliance produces most of the observed improvement, and computational adaptation, at least as implemented in this specific curriculum, does not yet contribute positively on top of it on average. This does not contradict the broader case made by Li *et al.* and colleagues that compliant morphology reduces the burden on control, nor Graule *et al.*'s finding that SAC performs well on soft-robot tasks; if anything, both are reinforced by Condition B's result. It does complicate the case, made across the curriculum learning literature reviewed in Section 2.4, that staged training reliably improves on end-to-end learning. At minimum, that benefit appears conditional on each stage's reward being well-posed relative to the skill it targets, a condition this study's Stage 4 did not satisfy. Because only one curriculum design was tested, this is a negative result about that design, not a general comparison between developmental and conventional RL.

Two limitations bound how far these conclusions should be extended. First, as detailed in Section 6, all results reflect only three random seeds per condition. While this is enough to reveal that run-to-run variance is real, substantial for Conditions A, B, and D, and in Condition B's case widening rather than narrowing as the third seed was added, it is too few to characterize the underlying distribution with confidence, and the reported standard deviations [Table 14] should be read as illustrative of the scale of variation rather than as precise estimates. Second, all four conditions were trained on a shared simulated platform with the modeling simplifications described in Section 6. The reaching and force-calibration difficulties identified here are specific to this simulation's contact model and action interface (raw joint-angle targets rather than task-space control, discussed in Section 3.5) and may not transfer directly to a physical system with different sensing noise, actuation dynamics, or contact fidelity. Section 6 also details the evaluation-protocol limitations (checkpoint decisions partly informed by test-object evaluations, training budgets not strictly matched across conditions, and asymmetric rigid and soft contact models) that further restrict these conclusions to directional, simulator-specific claims. These limitations are consistent with this project's stated scope: a controlled simulation study intended to isolate design factors before physical realization, with hardware validation and the associated sim-to-real questions left explicitly to the second study referenced in Section 7.

LIMITATIONS

Object-shape modeling. Our contact model does not capture real non-convex geometry. The irregular items (rock and toy) were simulated using lower friction and larger effective sizes rather than their actual shapes. This is a modeling simplification rather than an inherent property of the system.

Asymmetric contact models. The rigid and soft grippers are represented by two hand-designed contact models that differ in alignment tolerance, force distribution, and failure logic (Section 3.1). Because these parameters were not varied, part of the

morphology effect reported here may be built into the simulator's assumptions rather than arising from compliance in general. A sensitivity analysis over alignment tolerance, pad stiffness and damping, friction, force distribution, deformation limits, and crush thresholds is needed before the effect can be attributed to compliance beyond this implementation.

Checkpoint selection and the held-out test set. Several checkpoint-retention and stopping decisions reported in Section 4, most directly the restoration of Condition B seed 2's 600,000-step checkpoint, were informed by deterministic evaluations that included the nine test objects. For those runs, the reported generalization scores are therefore not fully independent held-out estimates and may be optimistically biased toward the retained checkpoint. A stricter protocol would separate a validation object set from a locked test set, freeze all extension, stopping, and checkpoint decisions using validation data only, and evaluate the test set once per seed; the study has not yet been rerun under that protocol.

Training budgets and stopping rules. The final four-way comparison is not strictly budget matched: Conditions A and B used 600,000 steps (with selected B seeds extended), while C and D used up to 1,500,000 and 1,100,000 steps with stage-specific stopping (Section 4.1). The final condition means therefore combine the effects of morphology and curriculum with those of compute, early stopping, and checkpoint selection. A primary, budget-matched factorial run with a predeclared stopping rule, with adaptive extensions reported separately as diagnostics, would give a cleaner causal separation.

Scope of the curriculum comparison. Only one developmental curriculum was evaluated, and no alternative curriculum or automated-progression baseline was run under the same action space, object set, budget, and evaluation protocol. The curriculum results

therefore describe this six-stage design specifically and should not be read as a general comparison between staged and end-to-end learning.

Action repeat as a methodological necessity. With the joint-angle action space and rate-limited PD control described in Section 3.5, no agent could be trained without control decimation. This is a substantive finding about sparse-reward manipulation RL with rate-limited low-level control, not an incidental implementation detail, and is reported as such rather than omitted.

Three seeds per condition, not a full distributional estimate. Every condition was trained with three independent random seeds (0, 1, and 2); results are reported as individual seed values, as mean ± sample standard deviation [Table 14], and as seed-matched effects [Table 15]. This is a meaningful improvement over a single-seed or two-seed report. Adding the third seed did not narrow the variance already visible with two; if anything, Condition B's spread widened (from 5 to 6 times across two seeds to 6.7 to 8.0 times across three), and Condition A's third seed reached a genuine floor of 0.0% on both metrics, a result the first two seeds alone did not anticipate. Three seeds remain too few to fit a reliable distribution or report informative confidence intervals, and the fact that variance grew rather than stabilized between two seeds and three is itself a caution against assuming three is enough either. Likewise, the 20 evaluation episodes per object are repeated trials on the same simulated object, not independent generalization samples, and the test set contains only nine object types. The direction of the morphology effect is unchanged in every seed-matched comparison; the curriculum effect on the soft gripper is not (on seed 2, D slightly exceeds B), and the exact magnitudes should be read with the variance in Table 14 in mind rather than as fixed point estimates. The gradually-ramped overshoot penalty proposed in Section 5 was tested on seed 0 of both grippers. It produced an actually holding-capable intermediate checkpoint on the soft gripper, though one that continued training subsequently degraded and that still fell short of simply stopping the curriculum at Stage 3, and no equivalent rescue on the rigid

gripper at the one intermediate checkpoint checked. A full re-run across all three seeds, and a systematic search for the best stopping point rather than checking only whichever checkpoints happened to be saved, was judged not to be a productive use of further compute given how narrow and timing-sensitive this result already appears; the underlying reward-schedule direction is not pursued further in this work.

A related, more specific caution concerns the still-climbing stopping rule itself, not just the seed count it was applied to. Section 4.3 documents that Condition B's seed 1 and seed 2 were both extended under an identical still-climbing criterion, with opposite results: seed 1's extension actually improved its result, while seed 2's extension regressed it, a regression only caught and corrected because a periodic checkpoint from before it happened was still available. Later seeds were held to their originally planned budgets without case-by-case extension, precisely to avoid compounding this inconsistency further. The rule that decides when a run is truly finished is therefore not fully reliable on its own, and any future work applying it should treat it as a useful heuristic for catching clearly under-trained runs, not as a guarantee that extending a still-climbing curve will help.

Sim-to-real transfer is out of scope for this paper. Per Section 2.5, transfer fidelity is not guaranteed to be equal across rigid and compliant hardware, and this work does not attempt to measure it. Physical hardware realization of the arm and both gripper variants is planned as a second, separate study, detailed next, that will report sim-to-real transfer explicitly and cite the present work.

FUTURE WORK

The immediate next phase of this project is physical hardware realization, planned as a separate study that will cite the present work as its simulation-side foundation, a deliberate two-paper split, common practice in robotics research, that lets each study be evaluated on its own terms. This paper's contribution is the controlled simulated

comparison and the reward-design finding in Section 5; the planned second paper's contribution is sim-to-real transfer validation, which raises its own distinct methodological questions (Section 2.5) better addressed on their own rather than folded into an already-large factorial study.

The planned physical platform is a 3D-printed, 4-degree-of-freedom arm matching the joint configuration of the simulated arm in this study (base yaw, shoulder pitch, elbow pitch, wrist pitch), preserving direct comparability between simulated and physical joint-space control rather than introducing a different kinematic structure at the hardware stage. The gripper end-effector will use an interchangeable mounting mechanism, allowing the rigid and soft (cephalopod-inspired) gripper variants to be swapped on the same physical arm rather than built as two separate robots. This preserves, in hardware, the same controlled-comparison philosophy applied throughout this paper: the arm, sensing, and control stack held constant while only the gripper varies, rather than confounding a sim-to-real comparison with two physically different platforms. Figure 4 summarizes the resulting hardware data flow, which mirrors the action/observation loop of the RL formulation in Section 3.3.

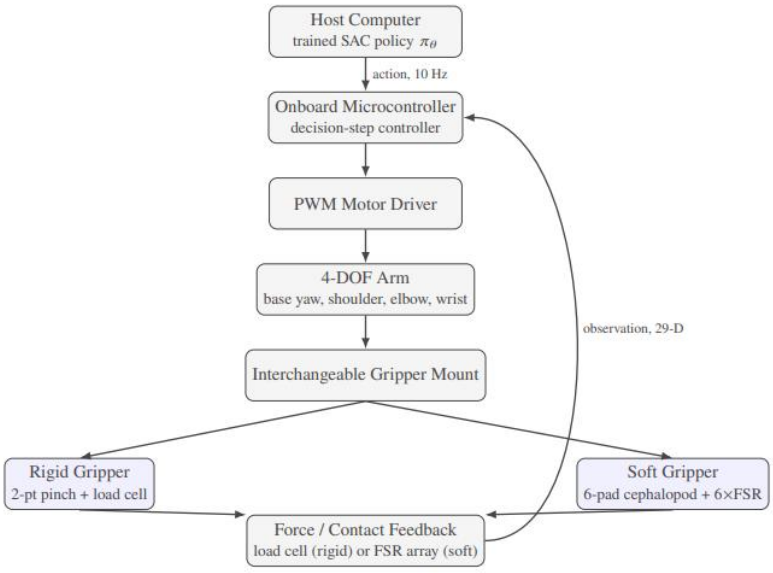


Figure 4. Planned hardware architecture for the follow-up study (Section 7). The action/observation loop mirrors the RL formulation of Section 3.3.

Three questions motivate this planned follow-up directly from the present study's findings. The first is whether Condition B's simulated advantage (Section 4.3, Figure 3), and its surprisingly large seed-to-seed variance (Section 4.6), persists under real sensing noise, actuation backlash, and imperfect contact modeling, or whether some of that advantage, or its variance, depends on simplifications specific to this simulation (Section 6). The second is whether the reward-design flaw identified in Sections 4.4 and 4.5 and discussed in Section 5 manifests identically on physical hardware, where force sensing and actuation are not perfectly precise the way a simulated gripper's commanded force is; a real gripper's own hardware noise could plausibly interact with an already-unstable reward landscape in ways this simulation cannot capture. This question is sharpened by the ramped-coefficient result in Section 5, which recovered only a narrow, timing-sensitive holding capability that continued training then degraded again, well short of a reliable fix on either gripper, and points toward the action-space interface, rather than the reward term's shape, as the more likely root cause; the planned hardware platform's motor and controller stack [Figure 4] gives a natural opportunity to test a task-space or IK-solved action interface directly, rather than carrying the raw joint-angle interface forward unexamined. The third is whether the interchangeable-gripper hardware mechanism itself introduces transfer artifacts, such as added compliance or backlash at the mounting interface, that a simulation assuming a rigid mount between arm and gripper does not account for. Reporting sim-to-real transfer as an explicit, measured result, rather than an assumed formality, for each of the four conditions studied here is the planned second paper's central contribution.

CONCLUSIONS

This paper presented Nazakat, a controlled 2 × 2 factorial study isolating the individual and combined contributions of compliant gripper morphology and

developmental-curriculum reinforcement learning to robust robotic manipulation, evaluated across three independent random seeds per condition. In this simulator, compliant morphology produced the most consistent effect, improving on the rigid gripper in every seed-matched comparison under both learning strategies, with its best single run (Condition B, seed 0) reaching 100.0% train and 85.0% generalization. The developmental curriculum, as currently implemented, did not improve on standard RL on average on either gripper, and never exceeded the rigid-gripper baseline on any seed. An uncapped force-overshoot penalty in its later stages, confirmed independently on both grippers and, within Condition C, on two of its three seeds (the third stopped before reaching that stage for an independent reason detailed in Section 4.4), taught the policy to avoid gripping rather than calibrate force, driving both curriculum conditions below their own demonstrated intermediate capability. The multi-seed comparison also surfaced a finding the single-seed draft of this paper could not have shown, and which the third seed strengthened rather than resolved: Condition B, this study's strongest result, is also its least reproducible, with a spread that widened from 5 to 6 times across two seeds to 6.7 to 8.0 times across three, a real tempering of how confidently morphology's advantage should be stated, even as its direction remains unchanged across all three runs. These conclusions are directional and simulator-specific: checkpoint decisions partly informed by test-object evaluations, unequal training budgets, and asymmetric contact models (Section 6) all limit how strongly they can be stated. The controlled research design is what surfaced the reward-design problem as a specific, fixable issue rather than a general indictment of staged learning. Physical prototyping as well as sim-to-real validation will remain for the envisioned second study in Section 7 that will build on and reference this work.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study, methodology, and formal analysis, and drafted the manuscript: Sharma P;

Developed the simulation and training software: Sharma P, Belsare V, Jaiswal R;

Performed data curation: Belsare V, Jaiswal R;

Performed validation: Sharma P, Dalvi R;

Performed investigation: Sharma P, Khare P;

Performed visualization: Belsare V;

Reviewed and edited the manuscript: Sharma P, Dalvi R, Khare P;;

Provided supervision: Sharma P, Tiwari M

Provided project administration: Sharma P;

Availability of data and materials

The simulator, training and evaluation scripts, rigid and soft contact-model parameters, object definitions, reward and stage-transition code, random seeds, evaluated checkpoints, checkpoint-selection and extension logs, dependency versions, and per-episode evaluation outcomes can be made available on request.

AI and AI-assisted tools statement

During the preparation of this work, the authors used Claude (Anthropic) to improve the language of the manuscript and to format it for submission. The authors reviewed and edited all AI-assisted content and take full responsibility for the published material.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. NASA Jet Propulsion Laboratory. Mars Exploration Rover Spirit. https://www.jpl.nasa.gov/missions/mars-exploration-rover-spirit-mer-spirit/ (accessed 2026-09-30).

2. McDonnell, C.; Arana-Catania, M.; Upadhyay, S. Autonomous robotic arm manipulation for planetary missions using causal machine learning. *In Proceedings of the 17th Symposium on Advanced Space Technologies in Robotics and Automation (ASTRA 2023),* Leiden, The Netherlands, 2023.

3. Nesnas, I. A. D.; Fesq, L. M.; Volpe, R. A. Autonomy for space robots: Past, present, and future. *Curr. Robot. Rep.* 2021, 2, 251-63.[DOI: 10.1007/s43154-021-00057-2]

4. Farag, H. O.; Gaber, M. M.; Awad, M. I.; Elhady, N. E. Myoelectric prosthetic hands: A review of muscle synergy, machine learning and edge computing. *ACM Comput. Surv.* 2025, 57, 311.[DOI: 10.1145/3742471]

5. Li, L.; Crosby, D.; Shuttleworth, M.; et al. A comparative analysis and scoping review of soft-rigid and industrial parallel rigid grippers. *Adv. Intell. Syst.* 2025, 7, 2400503.[DOI: 10.1002/aisy.202400503]

6. Wu, M.; Zheng, X.; Liu, R.; et al. Glowing sucker octopus (Staurothetus syrtensis)-inspired soft robotic gripper for underwater self-adaptive grasping and sensing. *Adv. Sci.* 2022, 9, 2104382.[DOI: 10.1002/advs.202104382]

7. Calisti, M.; Giorelli, M.; Levy, G.; et al. An octopus-bioinspired solution to movement and manipulation for soft robots. *Bioinspir. Biomim.* 2011, 6, 036002.[DOI: 10.1088/1748-3182/6/3/036002]

8. Flash, T.; Zullo, L. Biomechanics, motor control and dynamic models of the soft limbs of the octopus and other cephalopods. *J. Exp. Biol.* 2023, 226, jeb245295.[DOI: 10.1242/jeb.245295]

9. Choi, S.; Jang, J.; Lee, J.; Kim, D. W. Design and sensing frameworks of soft octopus-inspired grippers toward artificial intelligence. *Biomimetics* 2025, 10, 813.[DOI: 10.3390/biomimetics10120813]

10. Xie, Z.; et al. Octopus-inspired sensorized soft arm for environmental interaction. *Sci. Robot*. 2023, 8, eadh7852.[DOI: 10.1126/scirobotics.adh7852]

11. Wu, M.; Afridi, W. H.; Wu, J.; et al. Octopus-inspired underwater soft robotic gripper with crawling and swimming capabilities. *Research* 2024, 7, 0456.[DOI: 10.34133/research.0456]

12. Laschi, C.; Cianchetti, M.; Mazzolai, B.; Margheri, L.; Follador, M.; Dario, P. Soft robot arm inspired by the octopus. *Adv. Robot.* 2012, 26, 709-27.[DOI: 10.1163/156855312X626343]

13. Zhou, Z.; Zuo, R.; Ying, B.; et al. A sensory soft robotic gripper capable of learning-based object recognition and force-controlled grasping. *IEEE Trans. Autom. Sci. Eng.* 2024, 21, 844-54.[DOI: 10.1109/TASE.2022.3228255]

14. James, J. W.; Lepora, N. F. Slip detection for grasp stabilization with a multifingered tactile robot hand. *IEEE Trans. Robot.* 2021, 37, 506-19.[DOI: 10.1109/TRO.2020.3031245]

15. Han, D.; Mulyana, B.; Stankovic, V.; Cheng, S. A survey on deep reinforcement learning algorithms for robotic manipulation. *Sensors* 2023, 23, 3762.[DOI: 10.3390/s23073762]

16. Newbury, R.; et al. Deep learning approaches to grasp synthesis: A review. *IEEE Trans. Robot.* 2023, 39, 3994-4015.[DOI: 10.1109/TRO.2023.3280597]

17. Mahler, J.; Matl, M.; Satish, V.; et al. Learning ambidextrous robot grasping policies. *Sci. Robot.* 2019, 4, eaau4984.[DOI: 10.1126/scirobotics.aau4984]

18. Kalashnikov, D.; et al. Scalable deep reinforcement learning for vision-based robotic manipulation. *In Proceedings of the 2nd Conference on Robot Learning* (CoRL), Zürich, Switzerland, 2018; PMLR 87, pp. 651-73.[DOI: 10.48550/arXiv.1806.10293]

19. Zeng, A.; Song, S.; Welker, S.; Lee, J.; Rodriguez, A.; Funkhouser, T. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. *In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS), Madrid, Spain, 2018.[DOI: 10.1109/IROS.2018.8593986]

20. Joshi, S.; Kumra, S.; Sahin, F. Robotic grasping using deep reinforcement learning. *In Proceedings of the IEEE 16th International Conference on Automation Science and Engineering (CASE)*, Hong Kong, China, 2020.[DOI: 10.1109/CASE48305.2020.9216986]

21. Gupta, K.; Najjaran, H. Curriculum-based deep reinforcement learning for adaptive robotics: A mini-review. *Int. J. Robot. Eng.* 2021, 6.[DOI: 10.35840/2631-5106/4131]

22. Narvekar, S.; Peng, B.; Leonetti, M.; Sinapov, J.; Taylor, M. E.; Stone, P. Curriculum learning for reinforcement learning domains: A framework and survey. *J. Mach. Learn. Res.* 2020, 21, 1-50.[DOI: 10.48550/arXiv.2003.04960]

23. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. *In Proceedings of the 26th International Conference on Machine Learning (ICML),* Montreal, QC, Canada, 2009; pp. 41-8.[DOI: 10.1145/1553374.1553380]

24. Sun, Z.; Pang, B.; Yuan, X.; et al. Hierarchical reinforcement learning with curriculum demonstrations and goal-guided policies for sequential robotic manipulation. Eng. *Appl. Artif. Intell.* 2025, 153, 110866.[DOI: 10.1016/j.engappai.2025.110866]

25. Florensa, C.; Held, D.; Wulfmeier, M.; Zhang, M.; Abbeel, P. RREFERENCES 1. NASA Jet Propulsion Laboratory. Mars Exploration Rover Spirit. https://www.jpl.nasa.gov/missions/mars-exploration-rover-spirit-mer-spirit/ (accessed 2026-09-30).

2. McDonnell, C.; Arana-Catania, M.; Upadhyay, S. Autonomous robotic arm manipulation for planetary missions using causal machine learning. *In Proceedings of*

the 17th Symposium on Advanced Space Technologies in Robotics and Automation (ASTRA 2023), Leiden, The Netherlands, 2023.[DOI: 10.48550/arXiv.2403.00470]

3. Nesnas, I. A. D.; Fesq, L. M.; Volpe, R. A. Autonomy for space robots: Past, present, and future. Curr. Robot. Rep. 2021, 2, 251-63.[DOI: 10.1007/s43154-021-00057-2]

4. Farag, H. O.; Gaber, M. M.; Awad, M. I.; Elhady, N. E. Myoelectric prosthetic hands: A review of muscle synergy, machine learning and edge computing. *ACM Comput. Surv.* 2025, 57, 311.[DOI: 10.1145/3742471]

5. Li, L.; Crosby, D.; Shuttleworth, M.; et al. A comparative analysis and scoping review of soft-rigid and industrial parallel rigid grippers. *Adv. Intell. Syst.* 2025, 7, 2400503.[DOI: 10.1002/aisy.202400503]

6. Wu, M.; Zheng, X.; Liu, R.; et al. Glowing sucker octopus (Staurothetus syrtensis)-inspired soft robotic gripper for underwater self-adaptive grasping and sensing. *Adv. Sci*. 2022, 9, 2104382.[DOI: 10.1002/advs.202104382]

7. Calisti, M.; Giorelli, M.; Levy, G.; et al. An octopus-bioinspired solution to movement and manipulation for soft robots. *Bioinspir. Biomim.* 2011, 6, 036002.[DOI: 10.1088/1748-3182/6/3/036002]

8. Flash, T.; Zullo, L. Biomechanics, motor control and dynamic models of the soft limbs of the octopus and other cephalopods. *J. Exp. Biol.* 2023, 226, jeb245295.[DOI: 10.1242/jeb.245295]

9. Choi, S.; Jang, J.; Lee, J.; Kim, D. W. Design and sensing frameworks of soft octopus-inspired grippers toward artificial intelligence. *Biomimetics* 2025, 10, 813.[DOI: 10.3390/biomimetics10120813]

10. Xie, Z.; et al. Octopus-inspired sensorized soft arm for environmental interaction. *Sci. Robot.* 2023, 8, eadh7852.[DOI: 10.1126/scirobotics.adh7852]

11. Wu, M.; Afridi, W. H.; Wu, J.; et al. Octopus-inspired underwater soft robotic gripper with crawling and swimming capabilities. *Research* 2024, 7, 0456.[DOI: 10.34133/research.0456]

12. Laschi, C.; Cianchetti, M.; Mazzolai, B.; Margheri, L.; Follador, M.; Dario, P. Soft robot arm inspired by the octopus. *Adv. Robot.* 2012, 26, 709-27.[DOI: 10.1163/156855312X626343]

13. Zhou, Z.; Zuo, R.; Ying, B.; et al. A sensory soft robotic gripper capable of learning-based object recognition and force-controlled grasping. *IEEE Trans. Autom. Sci. Eng.* 2024, 21, 844-54.[DOI: 10.1109/TASE.2022.3228255]

14. James, J. W.; Lepora, N. F. Slip detection for grasp stabilization with a multifingered tactile robot hand. *IEEE Trans. Robot.* 2021, 37, 506-19.[DOI: 10.1109/TRO.2020.3031245]

15. Han, D.; Mulyana, B.; Stankovic, V.; Cheng, S. A survey on deep reinforcement learning algorithms for robotic manipulation. *Sensors* 2023, 23, 3762.[DOI: 10.3390/s23073762]

16. Newbury, R.; et al. Deep learning approaches to grasp synthesis: A review. *IEEE Trans. Robot.* 2023, 39, 3994-4015.[DOI: 10.1109/TRO.2023.3280597]

17. Mahler, J.; Matl, M.; Satish, V.; et al. Learning ambidextrous robot grasping policies. *Sci. Robot.* 2019, 4, eaau4984.[DOI: 10.1126/scirobotics.aau4984]

18. Kalashnikov, D.; et al. Scalable deep reinforcement learning for vision-based robotic manipulation. *In Proceedings of the 2nd Conference on Robot Learning (CoRL),* Zürich, Switzerland, 2018; PMLR 87, pp. 651-73.[DOI: 10.48550/arXiv.1806.10293]

19. Zeng, A.; Song, S.; Welker, S.; Lee, J.; Rodriguez, A.; Funkhouser, T. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. *In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* Madrid, Spain, 2018.[DOI: 10.1109/IROS.2018.8593986]

20. Joshi, S.; Kumra, S.; Sahin, F. Robotic grasping using deep reinforcement learning. *In Proceedings of the IEEE 16th International Conference on Automation Science and Engineering (CASE),* Hong Kong, China, 2020.[DOI: 10.1109/CASE48305.2020.9216986]

21. Gupta, K.; Najjaran, H. Curriculum-based deep reinforcement learning for adaptive robotics: A mini-review. *Int. J. Robot. Eng.* 2021, 6.[DOI: 10.35840/2631-5106/4131]

22. Narvekar, S.; Peng, B.; Leonetti, M.; Sinapov, J.; Taylor, M. E.; Stone, P. Curriculum learning for reinforcement learning domains: A framework and survey. *J. Mach. Learn. Res.* 2020, 21, 1-50.[DOI: arXiv.2003.04960]

23. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. *In Proceedings of the 26th International Conference on Machine Learning (ICML),* Montreal, QC, Canada, 2009; pp. 41-8.[DOI: 10.1145/1553374.1553380]

24. Sun, Z.; Pang, B.; Yuan, X.; et al. Hierarchical reinforcement learning with curriculum demonstrations and goal-guided policies for sequential robotic manipulation. *Eng. Appl. Artif. Intell.* 2025, 153, 110866.[DOI: 10.1016/j.engappai.2025.110866]

25. Florensa, C.; Held, D.; Wulfmeier, M.; Zhang, M.; Abbeel, P. Reverse curriculum generation for reinforcement learning. *In Proceedings of the 1st Annual Conference on Robot Learning (CoRL),* Mountain View, CA, USA, 2017; PMLR 78, pp. 482-95.[DOI: 10.48550/arXiv.1707.05300]

26. Wu, R.; Zhou, C.; Chao, F.; Zhu, Z.; Lin, C.-M.; Yang, L. A developmental learning approach of mobile manipulator via playing. *Front. Neurorobot.* 2017, 11, 53.[DOI: 10.3389/fnbot.2017.00053]

27. Savastano, P.; Nolfi, S. A robotic model of reaching and grasping development. IEEE Trans. *Auton. Ment. Dev.* 2013, 5, 326-36.[DOI: 10.1109/TAMD.2013.2264321]

28. Asada, M.; Hosoda, K.; Kuniyoshi, Y.; et al. Cognitive developmental robotics: A survey. *IEEE Trans. Auton. Ment. Dev.* 2009, 1, 12-34.[DOI: 10.1109/TAMD.2009.2021702]

29. Lungarella, M.; Metta, G.; Pfeifer, R.; Sandini, G. Developmental robotics: A survey. *Connect. Sci.* 2003, 15, 151-90.[DOI: 10.1080/09540090310001655110]

30. Oztop, E.; Bradley, N. S.; Arbib, M. A. Infant grasp learning: A computational model. *Exp. Brain Res*. 2004, 158, 480-503.[DOI: 10.1007/s00221-004-1914-1]

31. Law, J.; Lee, M.; Hulse, M.; Tomassetti, A. The infant development timeline and its application to robot shaping. *Adapt. Behav.* 2011, 19, 335-58.[DOI: 10.1177/1059712311419380]

32. Scheikl, P. M.; Tagliabue, E.; Gyenes, B.; et al. Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery. *IEEE Robot. Autom. Lett.* 2023, 8, 560-7.[DOI: 10.1109/LRA.2022.3227873]

33. Zhao, W.; Peña Queralta, J.; Westerlund, T. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. *In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI),* Canberra, Australia, 2020.[DOI: 10.1109/SSCI47803.2020.9308468]

34. Graule, M. A.; McCarthy, T. P.; Teeple, C. B.; Werfel, J.; Wood, R. J. SoMoGym: A toolkit for developing and evaluating controllers and reinforcement learning algorithms for soft robots. *IEEE Robot. Autom. Lett.* 2022, 7, 4071-8.[DOI: 10.1109/LRA.2022.3149580]

35. Williams, D. R. Mars Fact Sheet. NASA Goddard Space Flight Center, NSSDCA. https://nssdc.gsfc.nasa.gov/planetary/factsheet/marsfact.html (accessed 2026-09-30).

36. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *In Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018; PMLR 80, pp. 1856-65., *Mountain View,* CA, USA, 2017; PMLR 78, pp. 482-95.[DOI: 10.48550/arXiv.1801.01290]

26. Wu, R.; Zhou, C.; Chao, F.; Zhu, Z.; Lin, C.-M.; Yang, L. A developmental learning approach of mobile manipulator via playing. *Front. Neurorobot.* 2017, 11, 53.[DOI: 10.3389/fnbot.2017.00053]

27. Savastano, P.; Nolfi, S. A robotic model of reaching and grasping development. IEEE Trans. Auton. *Ment. Dev.* 2013, 5, 326-36.[DOI: 10.1109/TAMD.2013.2264321]

28. Asada, M.; Hosoda, K.; Kuniyoshi, Y.; et al. Cognitive developmental robotics: A survey. *IEEE Trans. Auton. Ment. Dev.* 2009, 1, 12-34.[DOI: 10.1109/TAMD.2009.2021702]

29. Lungarella, M.; Metta, G.; Pfeifer, R.; Sandini, G. Developmental robotics: A survey. *Connect. Sci.* 2003, 15, 151-90.[DOI: 10.1080/09540090310001655110]

30. Oztop, E.; Bradley, N. S.; Arbib, M. A. Infant grasp learning: A computational model. *Exp. Brain Res.* 2004, 158, 480-503.[DOI: 10.1007/s00221-004-1914-1]

31. Law, J.; Lee, M.; Hulse, M.; Tomassetti, A. The infant development timeline and its application to robot shaping. *Adapt. Behav.* 2011, 19, 335-58.[DOI: 10.1177/1059712311419380]

32. Scheikl, P. M.; Tagliabue, E.; Gyenes, B.; et al. Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery. IEEE Robot. *Autom. Lett.* 2023, 8, 560-7.[DOI: 10.1109/LRA.2022.3227873]

33. Zhao, W.; Peña Queralta, J.; Westerlund, T. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. *In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI),* Canberra, Australia, 2020.[DOI: 10.1109/SSCI47803.2020.9308468]

34. Graule, M. A.; McCarthy, T. P.; Teeple, C. B.; Werfel, J.; Wood, R. J. SoMoGym: A toolkit for developing and evaluating controllers and reinforcement learning algorithms for soft robots. *IEEE Robot. Autom. Lett.* 2022, 7, 4071-8.[DOI: 10.1109/LRA.2022.3149580]

35. Williams, D. R. Mars Fact Sheet. NASA Goddard Space Flight Center, NSSDCA. https://nssdc.gsfc.nasa.gov/planetary/factsheet/marsfact.html (accessed 2026-09-30).

36. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *In Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018; PMLR 80, pp. 1856-65.[DOI: 10.48550/arXiv.1801.01290]