electronics

MDPI

Article

Robust Curriculum-Based SAC for End-to-End Motion Control of a 7-DOF Manipulator Under Sparse Rewards

Yuhan Zhang and Jijun Gu *

College of Mechanical and Transportation Engineering, China University of Petroleum (Beijing), Beijing 102249, China; 2023010922@student.cup.edu.cn
* Correspondence: gu@cup.edu.cn

Abstract

End-to-end motion control of 7-degree-of-freedom (DOF) redundant manipulators under sparse reward signals presents a fundamental challenge in deep reinforcement learning (DRL) for robotics: the vast configuration space and absence of dense gradient information combine to produce severe cold-start failures and high cross-seed training variance. This paper proposes Curriculum-SAC-HER, a novel fusion framework integrating Soft Actor–Critic (SAC), Hindsight Experience Replay (HER), and a performance-driven three-stage Automatic Curriculum Learning (ACL) scheduler, designed to resolve the cold-start exploration bottleneck within a training budget of 300,000 environment interaction steps. The core methodology progressively expands the spatial target distribution across three stages of increasing difficulty, conditioning each stage transition on an 80% rolling success threshold to guarantee kinematic prior consolidation before advancing. A rigorous evaluation across 15 independent training runs (five seeds per group, all retained without filtering) demonstrates that the proposed framework achieves a final mean success rate of 84.8% (std: 11.0%), substantially surpassing the SAC + HER ablation (70.3%, Mann–Whitney U test, $p = 0.028$) and the DDPG baseline (22.3%, $p = 0.008$), while compressing cross-seed variance by 67% relative to the ablation. Zero-shot robustness evaluations under simulated domain perturbations further reveal that the learned policy maintains above 92% success across extreme friction variations and sustains 71.8% success under a 1.5× payload increase, demonstrating that the ACL module fosters generalized kinematic representations rather than over-fitting to specific contact mechanics.

Keywords: curriculum learning; soft actor–critic; hindsight experience replay; sparse rewards; redundant manipulator; robustness

Academic Editor: Flavio Canavero

Received: 13 May 2026
Revised: 17 June 2026
Accepted: 20 June 2026
Published: 24 June 2026

1. Introduction

Redundant manipulators are robotic systems possessing more degrees of freedom than strictly required to execute a given task, thereby enabling null-space motion—the ability to reconfigure the arm without altering the end-effector pose—which facilitates obstacle avoidance and joint-limit compliance during operation. The 7-DOF KUKA LBR iiwa arm employed in this study is a canonical example: since positioning and orienting an end-effector in three-dimensional space requires at most 6 DOF, the seventh degree of freedom constitutes a kinematic redundancy that endows the system with exceptional dexterity. This makes such manipulators highly valuable for complex tasks such as advanced manufacturing, surgical assistance, and dynamic obstacle avoidance. However, coordinating these high-dimensional systems requires solving highly nonlinear inverse

kinematics and dynamics problems. Traditional model-based control methods are computationally intensive and exhibit limited adaptability to unstructured or dynamically changing environments [1,2].

To overcome these limitations, Deep Reinforcement Learning (DRL) has emerged as a promising data-driven alternative, enabling robots to autonomously acquire complex motion skills through direct environmental interaction. In this context, end-to-end motion control—in which a policy maps directly from raw proprioceptive state observations to joint torque commands, bypassing explicit inverse kinematics or trajectory planning modules—represents a particularly demanding paradigm, as it requires the learning algorithm to simultaneously acquire kinematic reasoning and motor control from environmental interaction alone. Within this domain, maximum entropy frameworks—particularly the Soft Actor–Critic (SAC) algorithm [3]—have become the prevailing standard for continuous robotic control, owing to their favorable sample efficiency and principled exploration characteristics. Despite this considerable promise, transitioning DRL algorithms from simplified benchmark environments to kinematically complex simulated 7-DOF manipulators remains a formidable challenge. Current research frequently circumvents these physical complexities by employing artificially engineered dense reward functions, which demand extensive domain expertise and exhibit poor generalizability across heterogeneous task configurations [1,4].

Recent work has similarly documented that sparse reward signals in unstructured, high-dimensional environments severely impair policy generalization even under entropy-regularized training frameworks [5], motivating the need for structured exploration guidance. When standard SAC is applied directly to 7-DOF manipulators under sparse rewards, several critical vulnerabilities emerge. The inherent stochastic exploration mechanism provides no directional guidance within vast, high-dimensional state-action spaces, precipitating a severe cold-start problem: extreme sensitivity to initial random weight configurations and unacceptably high cross-seed training variance. A substantial fraction of training runs inevitably enter unrecoverable states—including uninformative value landscapes and catastrophic forgetting—resulting in complete convergence failure. This selective reporting convention—prevalent in benchmark-driven evaluations [1,2,4]—obscures the true distributional performance of DRL policies and complicates principled cross-method comparison.

To formalize this bottleneck, consider the geometric relationship between the operational workspace and the sparse success region. The global workspace volume ($x \in [0.5, 0.7]$, $y \in [-0.3, 0.3]$, $z \in [0.2, 0.5]$ m) amounts to 0.036 m^3, while the success region—a sphere of radius $\delta = 0.05$ m—occupies only $\approx 5.24 \times 10^{-4}$ m^3, yielding a hit probability of $P_{hit} \approx 1.45\%$ per end-effector position. Given a 500-timestep episode budget, the expected number of incidental successes under purely random exploration is approximately 0.73 per episode, rendering goal-proximal experience generation structurally unreliable without guidance. This analysis directly motivates the Stage-1 spatial constraint as a necessary architectural intervention rather than an arbitrary design choice.

Related Work on Curriculum Learning and Hindsight-Based Methods:

Several prior works have explored the integration of curriculum learning with hindsight-based goal relabeling to address sparse-reward challenges in robotic manipulation. Fang et al. [6] proposed Curriculum-guided Hindsight Experience Replay (CHER), which introduces a Goal-and-Curiosity-driven Curriculum to adaptively select which failed experiences are replayed, balancing proximity to the true goal and diversity of explored pseudo-goals across learning stages. Concurrently, Ren et al. [7] proposed Hindsight Goal Generation (HGG), an algorithmic framework that generates intermediate hindsight goals that are easy to achieve in the short term while remaining conducive to reaching the actual target in the long term, effectively constructing an implicit curriculum within

a fixed goal space. Both methods represent significant advances in sample efficiency for goal-conditioned sparse-reward reinforcement learning applied to robotic manipulation. Complementarily, Bing et al. [8] proposed graph-based hindsight goal generation (G-HGG), an extension of HGG applicable to obstacle-cluttered environments, demonstrating that graph-structured intermediate goal proposals substantially improve sample efficiency and collision avoidance in complex multi-step manipulation tasks.

However, CHER operates by selecting which experiences to replay within a fixed goal space (experience-selection curriculum), whereas the proposed framework progressively expands the spatial goal distribution itself (goal-space curriculum). These two mechanisms are orthogonal and complementary. The present work is therefore distinguished from CHER and HGG by its focus on resolving the cold-start problem through geometric task decomposition—a dimension not addressed by experience-selection or goal-generation strategies operating within a pre-defined workspace. More recently, Sayar et al. [9] proposed COHER, which couples HER with progressive environment shifts triggered by a success threshold; Bing et al. [10] introduced GC-HGG, which selects hindsight goals based on graph-based proximity and diversity metrics, enabling efficient learning in challenging manipulation tasks involving distant goals and obstacles. Beyond DRL-based approaches, Paolo et al. [11] proposed STAX, a quality-diversity algorithm that autonomously learns a behavior space on-the-fly and alternates between diverse policy exploration and reward exploitation via emitters, demonstrating competitive performance across sparse reward environments without requiring hand-designed behavior spaces. Or et al. [12] applied staged curriculum RL to high-DOF underactuated manipulators. These works operate within fixed goal spaces and do not address the cold-start bottleneck through geometric goal-space expansion as proposed here. These gaps collectively define the research opportunity addressed by this paper: a unified framework that resolves the cold-start problem at its geometric root, rather than managing its consequences through experience selection or goal relabeling within a pre-defined workspace.

To directly address the cold-start bottleneck identified above and the geometric limitations of existing methods, the primary objective of this research is not merely to report peak performance, but to systematically resolve the cold-start phenomenon and develop a sample-efficient, reproducible DRL framework for 7-DOF manipulator control. The core innovation lies in shifting the design paradigm from algorithmic hyperparameter tuning toward the provision of stable kinematic priors through structured task decomposition. Specifically, we propose a fusion architecture integrating progressive Automatic Curriculum Learning (ACL), SAC, and Hindsight Experience Replay (HER). The main contributions of this paper are threefold:

A Novel Robust Curriculum-Based SAC Framework for End-to-End Motion Control. To manage the exploration complexity of the 7-DOF configuration space within a training budget of 300,000 environment interaction steps, we design a three-stage geometric goal-space expansion strategy that progressively enlarges the spatial bounding box from which target goals are sampled. By conditioning stage transitions on an 80% rolling success rate threshold rather than fixed timestep intervals, this framework resolves the cold-start problem by guiding the agent from highly constrained proximal reaching tasks to the full unconstrained operational workspace.

Systematic Elimination of Training Variance and Failure Modes. We conduct a rigorous evaluation across five independent random seeds per group (15 total), retaining all initializations, including failures, to transparently quantify training variance. The proposed framework achieves a final mean success rate of 84.8% (std: 11.0%), substantially surpassing the SAC + HER ablation (70.3% mean, std: 33.6%) and the DDPG baseline (22.3% mean), while compressing cross-seed variance by 67%.

Zero-Shot Parametric Robustness and Asymmetric Sensitivity Analysis. We evaluate the learned policy under simulated domain perturbations without any fine-tuning or retraining. The results reveal strong friction invariance (above 92% success across extreme friction variations) and a physically interpretable asymmetric mass sensitivity (71.8% success under 1.5× payload increase), providing evidence that the ACL module develops generalized kinematic representations rather than memorizing specific contact dynamics.

The remainder of this paper is organized as follows: Section 2 presents the proposed Curriculum-SAC-HER framework, encompassing the SAC-HER learning substrate, the three-stage progressive curriculum scheduler, the direct observation strategy, and the reward function design. Section 3 describes the simulation setup and experimental evaluation, including comparative results against baseline methods, zero-shot robustness analysis, and failure mode discussion. Section 4 presents the conclusion and directions for future work.

2. Materials and Methods

2.1. Robotic Platform: KUKA LBR Iiwa Mechanical Configuration

The simulation platform employed in this study is based on the KUKA LBR iiwa 7-DOF manipulator, a serial-chain robotic arm comprising seven revolute joints arranged in a shoulder–elbow–wrist configuration (Figure 1). This architecture provides one degree of kinematic redundancy relative to the six degrees of freedom required for full end-effector pose specification in three-dimensional space, enabling null-space reconfiguration without disturbing the end-effector position. The arm has a nominal reach of approximately 0.8 m and a rated payload capacity of 7 kg. Under joint torque control—the actuation mode adopted in this study—the system dynamics are governed by the standard rigid-body equation of motion:

$$\tau = M(\theta)\ddot{\theta} + C\left(\theta, \dot{\theta}\right)\dot{\theta} + g(\theta)$$

where $\tau \in \mathbb{R}^7$ denotes the vector of applied joint torques, $M(\theta) \in \mathbb{R}^{7\times7}$ is the configuration-dependent inertia matrix, $C\left(\theta, \dot{\theta}\right) \in \mathbb{R}^{7\times7}$ captures Coriolis and centrifugal effects, and $g(\theta) \in \mathbb{R}^7$ is the gravitational torque vector. The high nonlinearity of M, C, and g across the 7-DOF configuration space renders precise analytical inversion computationally intensive and sensitive to model uncertainty—a fundamental motivation for the learning-based control approach developed in this paper. In the PyBullet simulation environment, the KUKA iiwa is represented via its official URDF model, with joint dynamics simulated at 240 Hz under the physics engine's rigid-body contact solver.

2.2. Soft Actor–Critic with Hindsight Experience Replay

To address the high-dimensional continuous control challenge of the 7-DOF manipulator, the foundational learning engine of the proposed framework relies on Soft Actor–Critic (SAC) [3], an off-policy deep reinforcement learning algorithm grounded in the maximum entropy framework. Unlike standard reinforcement learning formulations that exclusively maximize expected cumulative reward, SAC optimizes an augmented objective that incorporates the expected policy entropy:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} [r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))] \quad (1)$$

where ρ_π denotes the state-action marginals of the trajectory distribution induced by policy π, α is the temperature parameter governing the relative weight of the entropy term $\mathcal{H}$, and $r(s_t, a_t)$ is the reward function. This entropy maximization term systematically encourages broad exploration, which is theoretically critical for avoiding premature convergence in

high-dimensional continuous action spaces. The term 'deep' in this framework refers to the parameterization of both the policy π_ϕ and the twin value functions Q_{ψ_1}, Q_{ψ_2} as multi-layer neural networks—specifically, Multi-Layer Perceptrons (MLPs) with two hidden layers of 256 units each and ReLU activations. This neural network parameterization is what distinguishes SAC from tabular or linear RL methods: it enables function approximation across the continuous 20-dimensional state space $s_t \in \mathbb{R}^{20}$ and the 7-dimensional continuous action space $a_t \in [-1, 1]^7$, both of which are entirely intractable for non-parametric representations. The network architecture and full hyperparameter configuration are detailed in Section 2.5.

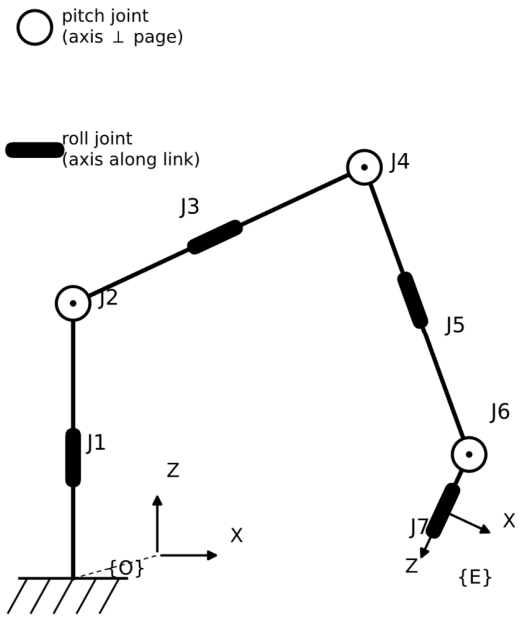


Figure 1. Kinematic schematic of the KUKA LBR iiwa 7-DOF manipulator, showing the seven joint axes (J1–J7) in the shoulder–elbow–wrist configuration, distinguishing pitch joints (axis perpendicular to the page) from roll joints (axis along the link), with the base frame {O} and end-effector frame {E} indicated.

However, in 7-DOF manipulation tasks characterized by strictly sparse rewards, standard SAC still faces a severe exploration bottleneck. When an agent fails to reach the target, the environment yields a constant step penalty, rendering collected transitions highly uninformative for neural network parameter updates. To systematically address this, Hindsight Experience Replay (HER) is integrated into the SAC architecture [13].

HER operates via retrospective goal relabeling. During environment interaction, the agent generates a trajectory of state transitions, actions, rewards, and the original goal, formally represented as a sequence of tuples $(s_t, a_t, r_t, s_{t+1}, g)$. We adopt the future relabeling strategy with a replay ratio of k = 4. For a transition at timestep t, the algorithm retrospectively samples k future states from the same episode (i.e., at timesteps $t' > t$) and designates them as virtual goals g'. The sparse reward function is then recomputed under each virtual goal to produce a substitute reward $r_t{}'$. The resulting hindsight-augmented transition tuple $(s_t, a_t, r'_t, s_{t+1}, g')$ is stored in the replay buffer alongside the original transition. This mechanism enables the agent to extract informative gradient signals from otherwise uninformative failure trajectories, substantially accelerating sample efficiency in sparse-reward settings.

Nevertheless, as the failure mode analysis in Section 3.4 will demonstrate, the SAC + HER combination alone is insufficient for reliable convergence in this setting. The vast dimensionality of the 7-DOF configuration space leaves the agent highly susceptible to

early-stage kinematic deadlocks and extreme sensitivity to initial weight configurations. Consequently, while SAC and HER constitute an essential learning substrate, they necessitate the integration of the proposed progressive Automatic Curriculum Learning (ACL) module to provide stable kinematic priors and facilitate reliable convergence within the constrained 300,000-timestep training budget. The multi-goal sparse-reward benchmark environments introduced by Plappert et al. [14] provide the canonical evaluation context for HER-based methods in robotic manipulation.

As illustrated in Figure 2, the framework comprises three interacting components. The Curriculum Scheduler maintains a sliding window of recent episode outcomes and triggers stage transitions upon reaching the 80% rolling success threshold, progressively expanding the spatial goal distribution from Stage 1 to Stage 3. The PyBullet Simulation Environment executes joint torque commands, applies domain randomization to friction and link mass parameters at each episode reset, and returns proprioceptive state observations alongside sparse reward signals. The SAC-HER Agent collects trajectories, performs hindsight goal relabeling with k = 4 future states, stores transitions in the replay buffer—which is retained across all stage transitions—and updates the actor and twin critic networks via entropy-augmented policy gradients.

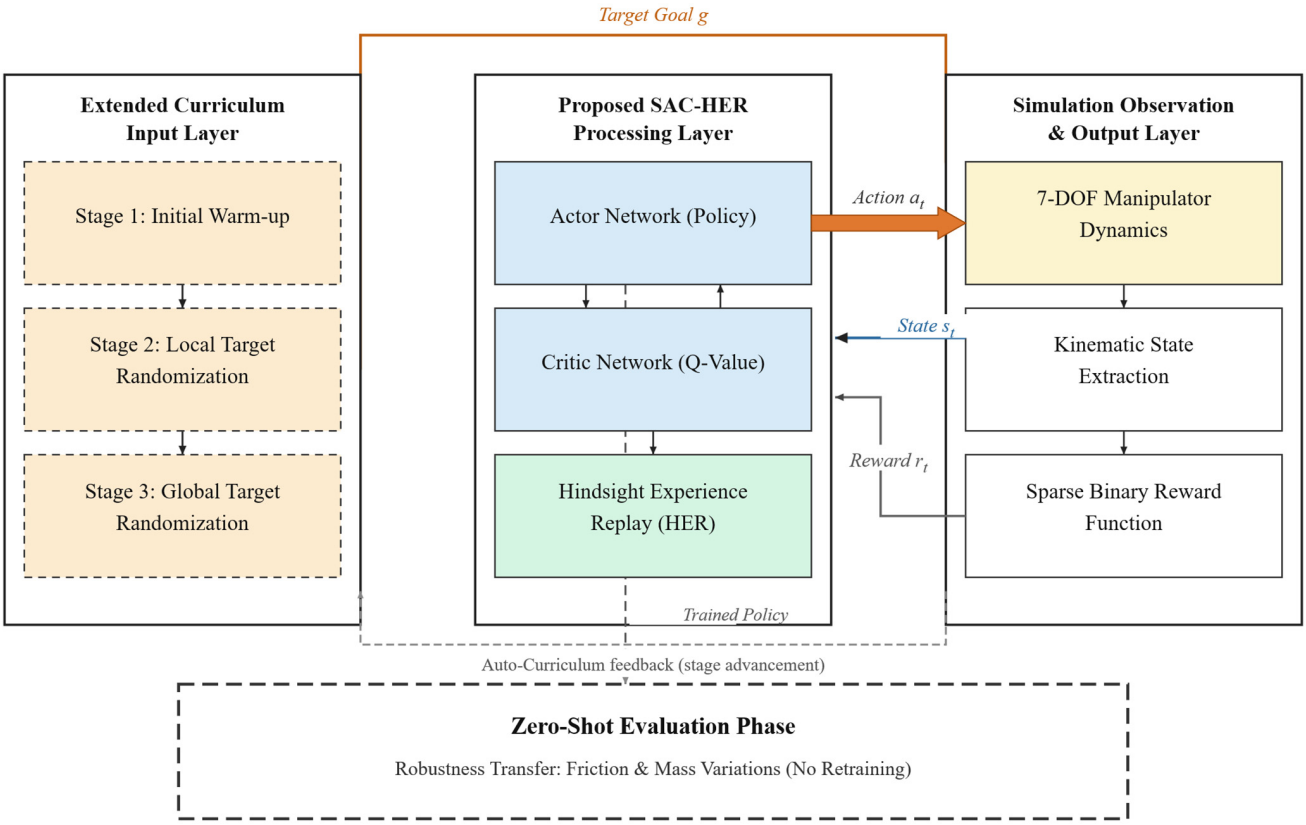


Figure 2. Architecture of the proposed Curriculum-SAC-HER framework, illustrating the data exchange among the Curriculum Scheduler, the PyBullet simulation environment, and the SAC-HER agent.

2.3. Core Innovation: Progressive Three-Stage Curriculum Scheduler

To systematically manage the high-dimensional exploration complexity of the 7-DOF manipulator and prevent the severe cold-start failures inherent in sparse-reward environments, this study proposes an Adaptive Automatic Curriculum Learning (ACL) framework. Curriculum learning for reinforcement learning—comprehensively surveyed by Narvekar et al. [15] as a principled framework for sequencing tasks or experiences of increasing diffi-

culty to accelerate policy acquisition—has been explored across a broad range of robotic control domains. Unlike prior curriculum approaches employing fixed, timestep-based stage transitions, the proposed scheduler employs a performance-driven dynamic mechanism. This design ensures that the agent fully consolidates the spatial kinematics of the current difficulty tier before advancing, thereby structurally suppressing catastrophic forgetting and stabilizing cross-seed training variance. This approach is referred to throughout as geometric goal-space expansion: the progressive enlargement of the uniform sampling bounding box $\mathcal{B}_k = \{g \mid g \sim \mathcal{U}(g_c - R_k,\ g_c + R_k)\}$ across stages $k \in \{1,\ 2,\ 3\}$, where the half-extent vector R_k grows monotonically with stage index. Unlike experience-selection methods such as CHER, which operate by filtering which transitions are replayed within a fixed goal distribution, geometric goal-space expansion modifies the goal distribution itself—directly reducing the volumetric mismatch between the agent's exploration range and the sparse success region.

2.3.1. Dynamic Transition Mechanism

The curriculum scheduler continuously monitors the agent's performance by maintaining a sliding window of the most recent $W = 100$ episodes. Let $S_i \in \{0, 1\}$ denote the binary success indicator of episode i. The moving-average success rate $\overline{S}$ is calculated as:

$$\overline{S} = \frac{1}{W} \sum_{i=1}^{W} S_i \tag{2}$$

A mandatory stage transition is triggered exclusively when $\overline{S} \geq 0.8$. The 80% threshold was selected based on two complementary design principles. First, a threshold substantially below 80% risks premature stage advancement before the agent has consolidated sufficient kinematic priors in the current stage, effectively reintroducing cold-start vulnerability at the next stage boundary. Second, a threshold approaching 100% would impose prohibitively long dwell times in early stages, consuming sample budget that would otherwise be available for the more challenging later stages. The value of 80% represents a practical balance between consolidation guarantee and sample efficiency, consistent with the convergence criterion convention adopted in related curriculum RL work on environment-shift-based sparse reward learning [9]. We acknowledge that a formal sensitivity analysis across alternative thresholds—for example, 60%, 70%, and 90%—would more rigorously characterize the robustness of this design choice; such ablation is deferred to future work, as it would require a substantial expansion of the simulation budget beyond the current 15 training runs. From a design perspective, the framework is expected to be relatively insensitive to a moderate threshold variation of $\pm 10\%$, since the performance-driven criterion functions primarily to ensure directional consolidation before difficulty increase, rather than requiring precision calibration of the exact threshold value. Upon triggering, the performance sliding window is reset and the spatial target distribution is expanded to the next stage's bounding box.

2.3.2. Replay Buffer Retention Policy Across Stage Transitions

A critical design decision concerns the treatment of the experience replay buffer at stage boundaries. In the proposed framework, the replay buffer is fully retained across all stage transitions; no flushing or experience filtering is performed. This decision is justified on the following grounds. Because SAC + HER is an off-policy algorithm, transitions collected under prior goal distributions remain valid training data provided that importance weighting is implicitly handled through the off-policy correction mechanism embedded in the Bellman update. Discarding the buffer at each transition would eliminate the kinematic priors accumulated during earlier, easier stages—precisely the structured knowledge that

the curriculum is designed to construct. Retaining prior transitions therefore serves as a form of implicit regularization: the critic continues to receive gradient signals from low-difficulty transitions, preventing the policy from catastrophically overwriting previously acquired motor primitives when confronted with the expanded goal distribution of the subsequent stage. Although the resulting replay buffer contains a non-stationary mixture distribution across stages, this heterogeneity is well-tolerated by the off-policy SAC update, as the critic's Bellman target is computed using the current policy rather than the behavior policy that generated the transitions. This design choice directly addresses the distributional shift concern intrinsic to staged curriculum learning and constitutes a principled mechanism for gradient stability across stage boundaries.The key design decisions of the proposed framework are summarized in Table 1.

Table 1. Summary of Key Design Decisions and Rationale for the Curriculum-SAC-HER Framework.

Design Decision	Choice	Justification
Buffer at stage transition	Retained (no flush)	Off-policy SAC tolerates non-stationary data; flushing destroys Stage 1–2 kinematic priors
Stage transition criterion	Performance-based ($\overline{S} \geq 0.8$)	Guarantees mastery before difficulty increase; prevents premature exposure to harder distributions
Performance monitor at transition	Reset (window cleared)	Ensures the threshold is re-earned from scratch at each new stage difficulty level
HER strategy	Future (k = 4)	Maximizes relabeling density from recent trajectory segments; standard best practice [13]
Entropy regularization across stages	Active throughout	SAC entropy term actively discourages freezing/deterministic degenerate policies at all stages

2.3.3. Progressive Spatial Expansion

In all stages, the target goal g is uniformly sampled from a 3D bounding box centered at a reachable reference coordinate $g_c = [0.6, 0.0, 0.35]^T$ (in meters). The curriculum scales the spatial half-extents $R_k = [r_x, r_y, r_z]^T$ for each stage k as:

$$g \sim \mathcal{U}(g_c - R_k, g_c + R_k) \tag{3}$$

The half-extent values R_1, R_2, and R_3 were determined through preliminary pilot simulations. Specifically, Stage 1 bounds were selected to ensure that undirected end-effector exploration produces an initial success rate exceeding 20% per episode, providing sufficient positive transitions to bootstrap HER. Stage 2 bounds were then set to approximately triple the Stage 1 volume, calibrated to retain a non-zero HER relabeling rate (above 10%) while meaningfully expanding spatial coverage. Stage 3 bounds were set to encompass the full designated operational workspace, as defined by the physical reachability limits of the KUKA LBR iiwa arm within the PyBullet simulation. No further tuning of these boundaries was performed after the initial pilot phase.

Stage 1 (Initial Kinematic Warm-up):

Spatial Bounds: $R_1 = [0.02, 0.05, 0.02]^T$.

Learning Objective: The target is constrained to a highly localized proximal region. This narrow target volume substantially elevates the probability of the randomly exploring end-effector intersecting the goal region, reliably generating the critical initial success signals required to bootstrap HER. Although the sparse success criterion ($\|p_{ee} - g\|_2 \leq 0.05$)

remains mathematically stringent, the geometric proximity of the goal to the agent's initial joint configuration ensures that even undirected stochastic exploration produces non-trivial positive transitions within the first episodes. This overcomes the cold-start bottleneck and populates the HER buffer with high-quality kinematic priors prior to any expansion of task difficulty. Critically, even within Stage 1, the target goal g is independently re-sampled at each episode reset from the uniform distribution over the Stage-1 bounding box (Equation (3)), ensuring that the agent generalizes across all positions within the constrained sub-space rather than memorizing a single fixed target. This per-episode re-sampling design structurally prevents fixed-target overfitting while preserving the geometric proximity guarantee required to bootstrap HER.

Stage 2 (Local Range Expansion):

Spatial Bounds: $R_2 = [0.05, 0.15, 0.075]^T$.

Learning Objective: Triggered after the agent reliably masters Stage 1, the target volume expands substantially. The primary objective at this stage is to exploit the retrospective goal-relabeling capability of HER. The agent leverages the foundational reaching trajectories consolidated in Stage 1 to explore this wider operational boundary, systematically generating a high density of near-miss experiences that HER converts into effective substitute gradient signals. This stage develops the agent's capacity for spatially generalized goal-conditioned reaching under moderate geometric uncertainty.

Stage 3 (Full Task Distribution):

Spatial Bounds: $R_3 = [0.10, 0.30, 0.15]^T$.

Learning Objective: Upon mastering Stage 2, the bounding box is expanded to encompass the full designated operational workspace. Policy optimization continues under these unconstrained conditions until the 300,000-timestep training budget is exhausted. This stage enforces global kinematic generalization across the full operational workspace, with target goals uniformly re-sampled from the global distribution at every episode—equivalent to an unconstrained random-target evaluation. The zero-shot parametric robustness results reported in Section 3.3 are therefore obtained under this full-workspace random-target condition, directly demonstrating the policy's generalization capability without any fixed-target overfitting.

2.3.4. Reward–Curriculum Interaction and Suppression of Freezing Behavior

A potential concern with applying a uniform -1 step penalty across all curriculum stages is that the agent, having learned conservative short-horizon policies in Stage 1's narrow workspace, may exhibit "over-cautious freezing behavior"—deliberately minimizing movement to limit accumulated penalties—when the bounding box expands abruptly at the Stage 3 transition. This concern is mitigated by two complementary mechanisms. First, the SAC entropy term actively penalizes deterministic, low-variance policies throughout training; a frozen arm constitutes a near-zero-entropy policy that the entropy regularizer directly discourages, regardless of the stage. Second, and more importantly, the retained replay buffer contains a dense distribution of successful trajectories from Stage 1 and Stage 2 that provide the critic with a rich baseline of positive-gradient transitions. Consequently, the value function entering Stage 3 is not a uniformly uninformative landscape—it already encodes a geometry-aware estimate of goal proximity. The agent therefore enters Stage 3 with a positively biased value prior that encourages continued goal-directed exploration rather than penalty minimization through inaction.The complete training procedure is summarized in Algorithm 1.

Algorithm 1: Curriculum-SAC-HER Training Procedure

Input: Total timestep budget $T_{\max}$ = 300,000
Curriculum stages $k \in \{1, 2, 3\}$ with bounds R_k
Sliding window size W = 100, threshold θ = 0.8
Replay buffer $\mathcal{B}$(capacity 500,000); retained across all stages
SAC networks: Actor π_ψ, Critics $Q_{\psi 1}$, $Q_{\psi 2}$, Target critics $Q_{\varphi 1'}$, $Q_{\varphi 2'}$
k_{HER} = 4, strategy: 'future'

Initialize: Stage k ← 1; success window $\{S_i\} \leftarrow \varnothing$; $\mathcal{B} \leftarrow \varnothing$
Randomly initialize all network weights φ, ψ1, ψ2

For each episode do:
1. Sample goal $g \sim \mathcal{U}(g_c - R_k, g_c + R_k)$
2. Reset environment; observe s_0
3. Collect trajectory $\tau = \{(s_t, a_t, r_t, s_{t+1})\}$ for $t = 0 \ldots T_{ep}$
   Where $a_t \sim \pi_\Phi(\cdot|s_t, g)$
4. [HER Relabeling]
   For each transition $(s_t, a_t, r_t, s_{t+1}, g)$ in τ:
   Store original transition in $\mathcal{B}$
   Sample k_{HER} future states $s_{t'}$, $t' > t$ from τ as virtual goals g'
   Recompute r_t = sparse(p_{ee}, g')
   Store hindsight transition $(s_t, a_t, r'_t, s_{t+1}, g')$ in $\mathcal{B}$
   ← NOTE: $\mathcal{B}$ is never flushed at stage boundaries
5. [SAC Update]
   Sample minibatch from $\mathcal{B}$
   Update Q_{ψ_1}, Q_{ψ_2} via Bellman residual minimization
   Update π_Φ via entropy-augmented policy gradient
   Soft-update target critics: $\psi' \leftarrow \tau_{soft} \cdot \psi + \left(1 - \tau_{soft}\right) \cdot \psi'$
6. [Curriculum Scheduler]
   Append episode outcome $S_i \in \{0, 1\}$ to sliding window
   Compute $\overline{S} = \frac{1}{W}\sum S_i$ over last W episodes
   If $\overline{S} \geq \theta$ AND $k < 3$:
   $k \leftarrow k + 1$ //Advance stage
   Clear sliding window//Reset performance monitor only
   //B is RETAINED//Preserve accumulated kinematic priors

Until total timesteps $\geq T_{\max}$
Output: Trained policy π_Φ

2.4. Direct Observation Strategy

To ensure full observability of the task dynamics, the agent employs direct state extraction via the PyBullet physics engine API. The continuous state space is formally defined as $s_t \in \mathbb{R}^{20}$. The state vector $s_t \in \mathbb{R}^{20}$ is constructed as follows:

$$s_t = [\ \underbrace{\theta_1, \ldots, \theta_7}_{\text{joint angles} \in \mathbb{R}^7}\ ,\ \underbrace{\dot{\theta}_1, \ldots, \dot{\theta}_7}_{\text{angular velocities} \in \mathbb{R}^7}\ ,\ \underbrace{p_{ee,x}, p_{ee,y}, p_{ee,z}}_{\text{EE position} \in \mathbb{R}^3},\ \underbrace{g_x, g_y, g_z}_{\text{goal position} \in \mathbb{R}^3}\ ] \tag{4}$$

The seven joint angles $\theta \in \mathbb{R}^7$ and angular velocities $\dot{\theta} \in \mathbb{R}^7$ are retrieved directly from the simulator to capture the complete instantaneous kinematic state of the manipulator. The end-effector Cartesian position $p_{ee} \in \mathbb{R}^3$ is computed via forward kinematics, and the goal position $g \in \mathbb{R}^3$ is provided by the curriculum scheduler at episode initialization. This compact proprioceptive representation deliberately excludes visual observations, enabling the agent to concentrate learning capacity on goal-conditioned trajectory planning without the additional overhead of image-based perception.

The SAC policy outputs a normalized continuous action vector $a_t \in [-1, 1]^7$. These normalized outputs are scaled by a uniform peak torque ceiling of 300 N·m to produce the final joint torque commands: $\tau_j = a_j \cdot 300$, $j = 1, \ldots, 7$. This uniform scaling is adopted for implementation simplicity within the simulation environment; the implications for Sim-to-Real transfer under heterogeneous actuator limits are acknowledged as a limitation in Section 3.4.3.

2.5. Reward Function Design and SAC Configuration

To ensure structural compatibility with the Hindsight Experience Replay (HER) mechanism and to eliminate the gradient distortion documented when dense distance-based penalties are combined with goal relabeling [13], we adopt a strictly sparse binary reward formulation that remains invariant across all three curriculum stages. The reward function r_t is defined as:

$$r_t = \begin{cases} 0, & if \|p_{ee} - g\|_2 \leq \delta \\ -1, & otherwise \end{cases} \tag{5}$$

where $\|p_{ee} - g\|_2$ denotes the Euclidean distance between the end-effector position and the target goal, and the success threshold δ is fixed at 0.05 m throughout training. Under this formulation, the agent receives a constant step penalty of -1 for every timestep in which the goal condition is not satisfied, and a terminal reward of 0 upon successful completion.

2.5.1. Interaction Between the Sparse Reward and Curriculum Stages

A potential concern with applying a uniform -1 step penalty across all curriculum stages is that the policy, having been optimized under the narrow, proximal target distribution of Stage 1, may develop over-cautious behavioral tendencies—specifically, minimizing cumulative penalty through reduced movement—that manifest as action freezing upon the expansion of the goal space at the Stage 3 transition. We argue that this concern is structurally mitigated by two complementary mechanisms intrinsic to the proposed framework.

First, the SAC maximum-entropy objective directly counteracts policy conservatism throughout all stages. The entropy regularization term $\alpha \mathcal{H}(\pi(\cdot \mid s_t))$ in the SAC objective penalizes low-variance, deterministic policies; a frozen or near-static arm constitutes a policy with near-zero entropy, which the entropy regularizer explicitly discourages irrespective of the current curriculum stage. Consequently, the temperature parameter α functions as a structural anti-freezing mechanism that remains active across stage transitions.

The replay buffer retention policy and its theoretical justification are detailed in Section 2.2; this design prevents catastrophic overwriting of Stage 1–2 kinematic priors at stage boundaries.

2.5.2. Network Architecture and Hyperparameters

The SAC agent employs twin critic networks $(Q_{\psi 1}, Q_{\psi 2})$ to mitigate value overestimation bias inherent in single-critic actor–critic methods. The policy (Actor) network and each of the two critic networks are implemented as Multi-Layer Perceptrons (MLPs). To accommodate the representational demands of the 7-DOF kinematic chain, network capacity is increased relative to standard low-dimensional control benchmarks: each net-

work comprises two hidden layers of 256 units with ReLU activations. ReLU activation functions are employed to facilitate gradient propagation through the deep network layers. Parameter optimization is performed using the Adam optimizer with a constant learning rate of 3×10^{-4}. The experience replay buffer capacity is set to 5×10^5 transitions, sufficient to retain the complete interaction history across the 300,000-timestep training lifecycle. A comprehensive listing of all hyperparameter settings is provided in Table 2.

Table 2. Key Hyperparameter Settings for the Curriculum-SAC + HER Model.

Hyperparameter	Value	Description
Algorithm	SAC + HER	Soft Actor–Critic (Off-Policy)
Optimizer	Adam	Adaptive Moment Estimation
Learning Rate	3×10^{-4}	Constant throughout training
Batch Size	256	Sampled from Replay Buffer
Replay Buffer Size	5×10^5	Experience storage capacity
Discount Factor (γ)	0.99	Long-term reward weighting
Soft Update	0.005	Target network update rate
Network Architecture	MLP [256, 256]	2 Hidden Layers, ReLU Activation
Curriculum Stages	3	Fixed → Local → Global
Max Steps per Episode	500	Timestep limit for a single trial
Total Training Timesteps	3×10^5	Total steps for full convergence
Simulation/Control Frequency	240 Hz/60 Hz (Action Repeat = 4)	Stepping rate for the physics engine and the high-level policy.
Max Torque Output	300 N·m	Physical upper limit of the joint actuators for the KUKA IIWA.
HER Goal Selection Strategy	‘future’ (k = 4)	Hindsight relabeling strategy using 4 sampled future states.

3. Simulation Evaluation and Failure Mode Analysis

3.1. Simulation Setup

Simulation Platform and Robot Dynamics: All simulations were conducted within the PyBullet physics engine, utilizing the KUKA IIWA (7-DOF) URDF model as the primary kinematic platform. To ensure high-fidelity physical interactions without excessive computational overhead, the base simulation stepping frequency was set to 240 Hz, while the high-level DRL control policy operated at 60 Hz (implementing an action repeat of 4). The manipulator was governed under joint torque control mode. To improve training robustness, mild domain randomization is applied at each episode reset: joint lateral friction is sampled uniformly from [0.5, 1.2] and link masses are scaled by a factor sampled uniformly from [0.9, 1.1], implemented via the PyBullet dynamics API. The SAC agent outputs normalized continuous actions, which are subsequently scaled by a uniform peak torque ceiling of 300 N·m to produce the final joint torque commands. Within each action repeat block, the same torque command is applied across all four simulation substeps; the sparse reward signal is computed once from the terminal state of the block rather than accumulated across substeps, ensuring reward semantics remain consistent with the single-step formulation.

Task Definition and Success Criteria: The core environment is formulated as an end-to-end 3D spatial reaching task. The target goal g is visually represented by a geometric sphere with a radius of 0.05 m. While the target’s spatial distribution is dynamically governed by the Three-Stage Curriculum Strategy (as detailed in Section 2.3), the ultimate operational workspace requires the agent to reach targets stochastically spawned within the global boundaries: $x \in [0.5,\ 0.7]$, $y \in [-0.3,\ 0.3]$, and $z \in [0.2,\ 0.5]$ (measured in meters).

An interaction episode is strictly defined as successful if the Euclidean distance d between the robot's end-effector position (p_{ee}) and the target center (g) satisfies $d \leq 0.05$ m. Upon reaching this precise spatial threshold, the environment yields a positive success flag and the episode terminates immediately. Conversely, to prevent infinite loops during early-stage exploration, an episode is forcibly truncated if the maximum limit of 500 timesteps is exceeded without meeting the distance requirement.

Computing Infrastructure: All simulations were conducted on a workstation running Windows 11, equipped with a 13th Gen Intel Core i7-1360P CPU (2.20 GHz) and 16 GB RAM. Training was executed on the CPU, as no discrete GPU was available in the simulation environment. The software stack comprised Python 3.12.7, PyTorch 2.3.1, Stable-Baselines 3 2.0, Gymnasium 0.29, and PyBullet 3.2.5. Each seed required approximately 2–6 h of wall-clock training time. No data preprocessing was applied, as all training data were generated online through agent–environment interaction within the PyBullet simulator; the statistical analyses reported in this study were performed directly on the episodic evaluation logs recorded during training. The KUKA IIWA platform and PyBullet environment adopted here are consistent with recent DRL studies on high-DOF manipulation control [10,14], including prior SAC + HER formulations for multi-arm path planning under sparse rewards [16] and DRL-based collision-avoidance trajectory planning for 7-DOF manipulators under uncertain environmental constraints [17], both of which confirm the viability of PyBullet simulation as an evaluation platform for complex continuous manipulation policies.

3.2. Baselines and Training Parameters

To rigorously validate the necessity of the proposed Curriculum-SAC-HER framework and systematically isolate the contributions of its key algorithmic components, we conduct a comparative study comprising three strictly controlled experimental groups. Each group is evaluated across five independent random seeds to quantify training variance and assess algorithmic stability. To ensure complete simulation transparency, all 15 training runs are retained in full without exception—no seed filtering, early termination, or selective reporting is applied at any stage of the analysis. Seeds that exhibit convergence failure or catastrophic forgetting are explicitly included in both the learning curve visualization and all reported statistics, as these failure modes constitute the primary empirical evidence motivating the proposed framework.

Proposed Method (Curriculum-SAC + HER): The complete fusion framework proposed in this work integrates the Soft Actor–Critic algorithm with Hindsight Experience Replay and is driven by the dynamic Three-Stage Automatic Curriculum Learning (ACL) scheduler, which progressively expands the target distribution based on a rolling 80% success-rate threshold.

Ablation Baseline (Standard SAC + HER): To isolate and quantify the contribution of the ACL module, we introduce a standard SAC agent equipped with HER but initialized and trained directly in the fully unconstrained global workspace—equivalent to Stage 3 of the curriculum—without any staged task decomposition. This baseline is specifically designed to expose the cold-start failure modes and extreme cross-seed variance that arise from unguided entropy-regularized exploration in the vast 7-DOF configuration space.

Algorithm Baseline (Standard DDPG—no HER, no Curriculum): To establish the fundamental difficulty of the sparse-reward reaching task, we introduce a standard Deep Deterministic Policy Gradient (DDPG) agent, trained without HER and without curriculum assistance, in the same unconstrained global environment as the ablation baseline. This configuration demonstrates that classical deterministic policy gradient methods relying on additive exploration noise are structurally insufficient to overcome the sparse-reward bottleneck in high-dimensional continuous manipulation tasks. It should be noted that

the DDPG baseline is included not to challenge state-of-the-art methods, but for historical and architectural completeness. The SAC + HER ablation already represents the strongest non-curriculum baseline by incorporating both maximum-entropy exploration and hindsight relabeling; DDPG therefore serves to illustrate the compounded benefit of all three components relative to a classical deterministic policy gradient method and to confirm that the sparse-reward bottleneck cannot be overcome by exploration noise alone.

Evaluation Metrics: To comprehensively assess learning efficiency, convergence reliability, and policy robustness, the following metrics are continuously monitored across all experimental groups:

Task Success Rate (S_{eval}): The primary performance metric, defined as the proportion of evaluation episodes in which the end-effector reaches the target position within the 0.05 m Euclidean distance threshold. Evaluation is conducted every 10,000 environment timesteps, during which training is paused and the agent executes 30 deterministic evaluation episodes. Crucially, all evaluation episodes are conducted in deterministic mode: the actor network outputs the mean of the policy distribution $\mu_\Phi(s_t)$ directly, with the stochastic sampling component disabled. This ensures that reported success rates reflect the learned policy's exploitative performance rather than its stochastic exploratory behavior during training.

Average Episodic Reward: Under the strictly sparse binary reward design, the cumulative episodic reward is a monotonic function of the number of successful timesteps within an episode, providing a complementary signal to the discrete success rate metric.

Convergence Speed and Variance: Convergence speed is operationally defined as the number of environment timesteps elapsed before the per-seed success rate first achieves and subsequently remains above 80% across three consecutive evaluation checkpoints (spanning 30,000 environment timesteps), as determined by post hoc inspection of the recorded evaluation logs. Seeds that do not satisfy this criterion within the 300,000-step budget are recorded as not converged. Cross-seed variance is reported as the standard deviation of the final success rate across all five seeds per group, computed over the last five evaluation checkpoints to reduce noise from transient fluctuations. It should be noted that this convergence criterion is operationally distinct from the curriculum advancement threshold described in Section 2.3: the curriculum advancement criterion evaluates an 80% rolling success rate over a 100-episode sliding window during training interactions, whereas the convergence metric above is applied to the periodic deterministic evaluation protocol conducted every 10,000 environment timesteps on the global Stage-3 task distribution. These two thresholds share the same numerical value but measure fundamentally different quantities at different timescales.

Statistical Significance. Given the small per-group sample size (n = 5), we supplement the descriptive statistics with pairwise Mann–Whitney U tests (one-tailed) to confirm that the observed performance differences are not attributable to random seed variation. The comparison between Curriculum-SAC-HER and Standard SAC + HER yields p = 0.028; the comparison with the DDPG baseline yields p = 0.008. These results meet the conventional α = 0.05 threshold; however, given the small sample size (N = 5), these p-values should be interpreted as indicative of superiority rather than definitive statistical proof, as discussed in Section 3.4.3.

Training Parameters and Implementation Details: To ensure a rigorously fair comparison, all experimental groups share identical core environment parameters and matched foundational hyperparameters wherever algorithmically applicable. The experience replay buffer capacity is set to 5×10^5 transitions for all off-policy algorithms, sufficient to encompass the complete 300,000-step training lifecycle. HER employs the standard future

goal selection strategy, with k = 4 relabeled goals per trajectory. Complete hyperparameter settings are provided in Table 2.

3.3. Simulation Results

This section presents the simulation results for all three evaluated methods. The results are organized as follows: Section 3.3 first reports comparative training performance across the three groups, followed by a quantitative summary in Table 3. Section 3.4 then provides mechanistic analysis of the observed learning dynamics, failure mode diagnosis, and zero-shot robustness evaluation. Together, these two sections correspond to the simulation results and interpretation dimensions that the structured experimental design in Sections 3.1 and 3.2 was designed to assess.

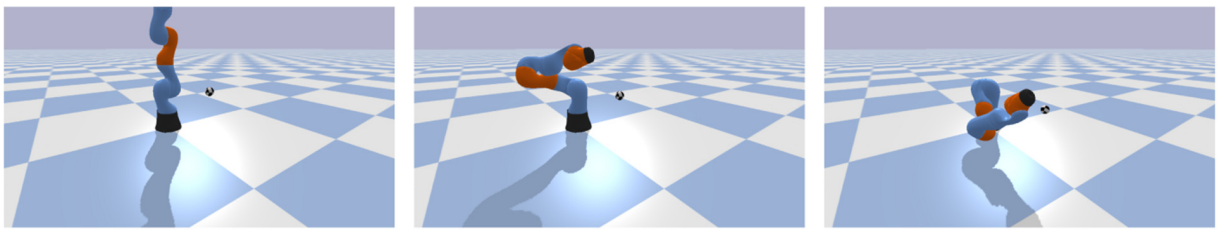
Table 3. Quantitative performance comparison of different methods evaluated across 5 independent random seeds under a 300,000-timestep budget.

Algorithm	Final Success Rate	Convergence Steps	Average Return
Ours (Curriculum-SAC-HER)	84.8 ± 11.0	~205,000 (4/5 seeds)	−20.3 ± 6.5
Ablation (Standard SAC + HER)	70.3 ± 33.6	~260,000	−25.4 ± 14.2
Baseline (standard DDPG)	22.3 ± 17.1	>300,000	−42.2 ± 3.8

Random seeds per experimental group: Curriculum-SAC-HER: {401, 402, 403, 404, 405}; Standard SAC + HER (Ablation): {201, 202, 203, 204, 205}; Standard DDPG: {301, 302, 303, 304, 305}.

The training-time mean success rate of 84.8% reflects the average of the last five periodic evaluation checkpoints (each computed over 30 deterministic episodes at 10,000-timestep intervals) across five independent seeds. The zero-shot baseline condition reports 94.0% (±9.5%), evaluated over 100 dedicated deterministic episodes per seed using the best-performing saved model checkpoint (best_model.zip) after training completion. The performance gap is attributable to two factors: the difference in evaluation sample size (30 vs. 100 episodes per checkpoint), and the distinction between the time-averaged training metric and the peak-performance best model, both of which are standard reporting conventions in deep reinforcement learning.

The simulation task execution sequence is illustrated in Figure 3. Figure 4 presents the learning curves of all three experimental groups over the full 300,000-timestep evaluation budget. The final performance comparison across all three groups is shown in Figure 5, and the quantitative summary is provided in Table 3.



(a) Initial State (b) Reaching Maneuver (c) Successful Grasp

Figure 3. Visualization of the simulation task execution sequence: (**a**) initial configuration of the 7-DOF KUKA LBR iiwa manipulator at episode reset; (**b**) reaching maneuver toward the randomized target position (red sphere, radius 0.05 m); (**c**) successful task completion upon the end-effector tip entering the target radius. Note that the present study frames the task as a spatial reaching problem rather than object grasping; no physical gripper is mounted, and the end-effector is represented by the terminal link's tip. Task success is defined solely by the Euclidean distance between the end-effector tip and the target center falling below the 0.05 m threshold.

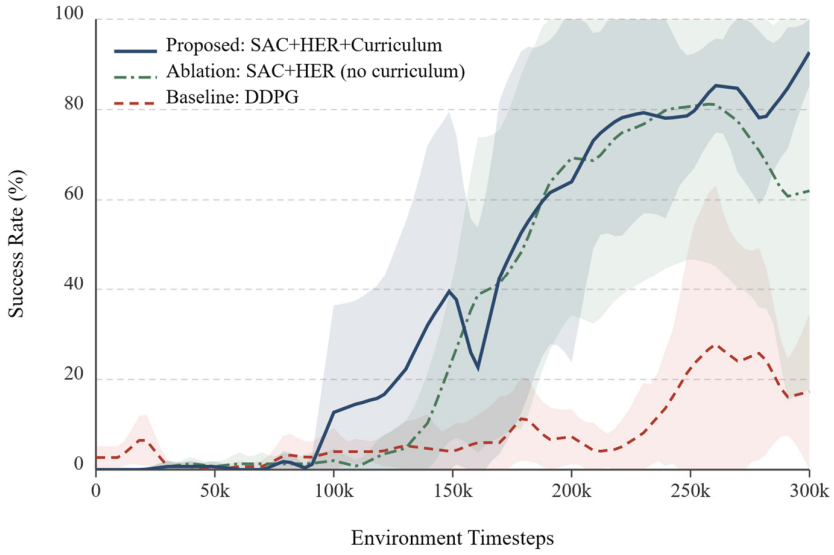


Figure 4. Learning curves of all evaluated methods over 300,000 environment timesteps (5 independent seeds per group; all seeds retained, including failures). The proposed Curriculum-SAC-HER (blue) achieves consistent convergence across all five seeds. The SAC + HER ablation (grey) exhibits extreme divergence: two seeds fail catastrophically (seeds 203 and 201, reaching final success rates of 9.3% and 58.7%, respectively), while three converge successfully—directly demonstrating the cold-start vulnerability of unguided sparse-reward training. The DDPG baseline (red) fails to acquire meaningful goal-directed behavior across the converging seeds. Shaded regions denote one standard deviation across seeds.

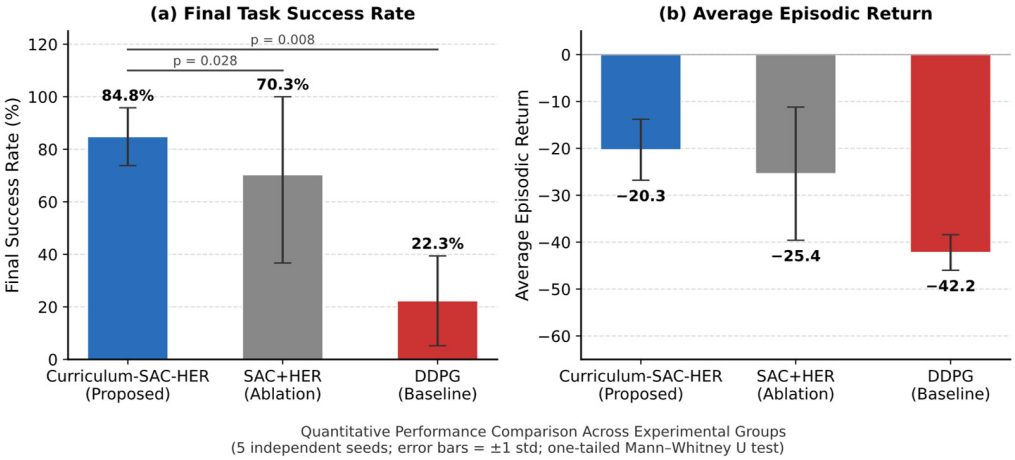


Figure 5. Final performance comparison across three experimental groups (n = 5 seeds; error bars = ± 1 std). Standard DDPG without HER and without curriculum assistance. Convergence defined as the first timestep at which the per-seed success rate stably exceeds the 80% threshold. Wilcoxon rank-sum test (one-tailed, n = 5): Curriculum-SAC-HER vs. Standard SAC + HER, $p < 0.05$; vs. Standard DDPG, $p < 0.05$. Having established the training-time superiority of the proposed framework, we next examine whether the learned policy generalizes beyond the training distribution—specifically, whether the kinematic priors instilled by the curriculum confer robustness to unmodeled physical perturbations encountered during zero-shot deployment, as shown in Figure 6.

Curriculum-SAC-HER. All five seeds converge within the training budget, with four of five satisfying the three-consecutive-checkpoint stability criterion; the remaining seed (Seed 405) reaches the 80% threshold only at the terminal evaluation checkpoint. A decisive acceleration in learning becomes apparent at approximately 100,000 timesteps, at which point the blue curve initiates a sustained breakout as the curriculum transitions the agent into the global task distribution of Stage 3. The method ultimately converges to a mean

final success rate of 84.8% (std: 11.0%), with four of five seeds exceeding 83% and reaching peak performance of 100% during training. Convergence to the 80% threshold is achieved at a mean of approximately 205,000 environment timesteps across the four converging seeds (bootstrap 95% CI: 160,000–252,500 steps), representing a mean sample complexity reduction of approximately 21.2% relative to the Standard SAC + HER ablation (bootstrap 95% CI: 2.9–38.5%). One seed (Seed 405) exhibited delayed convergence, reaching 80% only at the final evaluation checkpoint without satisfying the three-consecutive-checkpoint stability criterion within the 300,000-step budget; the stage-wise mechanistic basis for this efficiency gain is analyzed in Section 3.4.1. The residual cross-seed variance of ±11.0% reflects a combination of moderate variation in Stage 3 convergence speed across four converging seeds, and one seed (Seed 405) that reached the 80% threshold only at the terminal evaluation checkpoint without satisfying the three-consecutive-checkpoint stability criterion. Critically, no seed exhibits an uninformative value landscape or irrecoverable policy collapse of the kind observed in the ablation group.

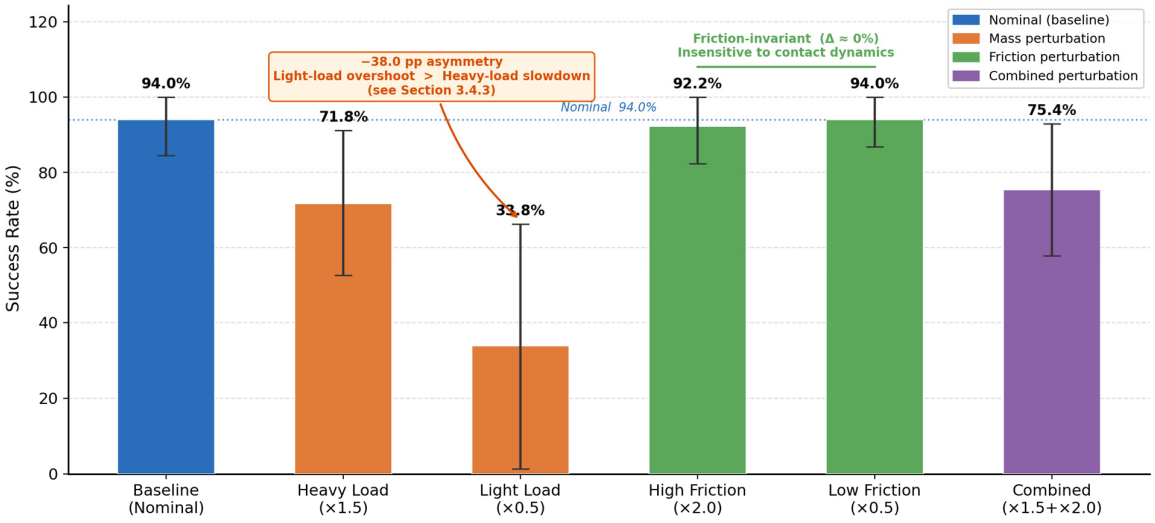


Figure 6. The parametric robustness evaluation results of the trained Curriculum-SAC-HER policy across six simulated perturbation conditions.

Standard SAC + HER (Ablation). The Standard SAC + HER ablation without curriculum guidance (dash–dot grey curve) produces a mean success rate of 70.3% (std: 33.6%). While three seeds eventually converge to high performance (above 90%), two seeds fail entirely—one stagnating permanently at 9.3% due to an uninformative value landscape under sparse rewards, and another collapsing to 58.7% following catastrophic forgetting. The resulting confidence band spans nearly the full success-rate range, from near 0% to near 100% across seeds, and is diagnostic of a policy acquisition process governed by initialization stochasticity rather than a reliable learning mechanism. A detailed failure mode analysis of these specific seed failures is provided in Section 3.4.2. This instability is a structural consequence of the cold-start problem: in the unconstrained global workspace, whether a seed converges is determined almost entirely by whether its initial weight configuration generates sufficient exploratory variance near the goal region to produce accidental early successes. Seeds that satisfy this condition escape the cold-start trap and eventually achieve high performance; those that do not enter an uninformative value landscape from which recovery within the 300,000-timestep budget is essentially impossible. The ACL module resolves this structurally by guaranteeing goal-proximal trajectories from the first episode, decoupling convergence from initialization luck entirely.

Standard DDPG (no HER, no Curriculum). The Standard DDPG baseline (dashed red curve) fails to acquire meaningful goal-directed behavior across all five seeds, yielding a mean final success rate of 22.3% (std: 17.1%). No seed achieves the 80% convergence threshold within the 300,000-step budget, confirming the structural inadequacy of deterministic exploration under sparse rewards in high-dimensional continuous control.

The quantitative summary is provided in Table 3. The final success rate improvements of Curriculum-SAC-HER over both baselines are statistically significant: a Wilcoxon rank-sum test (one-tailed, $n = 5$ per group) yields $p < 0.05$ for both pairwise comparisons, confirming that the observed performance differences are not attributable to random seed variation.

All evaluations are conducted without any fine-tuning or retraining of the policy, and results are reported as the mean and standard deviation over 100 deterministic evaluation episodes per condition across five independent seeds. Under the nominal baseline condition, the policy achieves a success rate of 94.0% ($\pm$9.5%), confirming that the trained policy retains strong task performance at inference time—notably higher than the training-time mean of 84.8%, as detailed in the opening paragraph of this section.

Under friction perturbations, the policy demonstrates strong invariance across both directions of modification. A high-friction condition yields a success rate of 92.2% ($\pm$9.8%), while a low-friction condition yields 94.0% ($\pm$7.2%). The near-complete preservation of performance across both friction perturbations indicates that the learned policy is largely insensitive to contact dynamic variations at the manipulator's joints within the tested range. It should be noted that this friction invariance is partially attributable to the mild friction domain randomization applied during training (lateral friction sampled from [0.5, 1.2]), rather than being solely a product of the curriculum-induced kinematic generalization. The policy's maintained performance under friction conditions exceeding the training randomization range is attributable to two mechanisms beyond domain randomization itself. First, the learned policy is fundamentally a kinematic reaching strategy: the SAC agent outputs joint torque commands calibrated to produce goal-directed end-effector displacements, and the sensitivity of this strategy to joint friction depends on whether friction variations produce large-magnitude deviations in end-effector trajectories. Within the tested perturbation range, joint-level friction primarily affects the damping characteristics of individual joints rather than substantially redirecting end-effector motion—an effect that attenuates sensitivity even for out-of-distribution friction values. Second, the SAC maximum-entropy objective explicitly penalizes low-entropy, brittle policies throughout training, encouraging the agent to develop action policies with broader stability margins than a purely reward-maximizing objective would achieve. This entropy regularization produces a degree of implicit robustness that extends beyond the nominal training distribution. Together, these two mechanisms explain why friction invariance generalizes beyond the [0.5, 1.2] training range, and why inertial perturbations—which directly alter the system's dynamic response to the learned torque commands at the level of joint accelerations—produce the qualitatively distinct asymmetric performance degradation described above.

Under mass perturbations, a pronounced asymmetry emerges between the two perturbation directions. The heavy-load condition ($\times$1.5 nominal payload) produces a moderate but graceful degradation to 71.8% ($\pm$19.4%). In contrast, the light-load condition ($\times$0.5 nominal payload) results in a substantially larger performance drop to 33.8% ($\pm$32.5%). A detailed control-theory mechanistic basis for this asymmetric sensitivity to payload perturbations is analyzed in Section 3.4.3.

Under the combined perturbation condition—in which both payload ($\times$1.5) and friction ($\times$2.0) are simultaneously perturbed—the policy achieves a success rate of 75.4%

(±17.5%). This result is notably higher than the light-load-only condition, suggesting a partial compensatory interaction between the two perturbation types under the combined condition.

Taken together, the zero-shot robustness results indicate that Curriculum-SAC-HER develops parametric generalization within the simulated environment beyond the nominal training distribution, exhibiting particularly strong invariance to friction perturbations while displaying a directionally asymmetric sensitivity to inertial perturbations.

3.4. Simulation Analysis and Discussion

Building upon the quantitative results presented above, this section provides an in-depth analysis of the underlying mechanisms driving the observed performance disparities. We explicitly discuss the synergistic roles of our Three-Stage Curriculum and Hindsight Experience Replay (HER) in overcoming sparse rewards, analyze the specific failure modes of the baseline algorithms in high-dimensional continuous control, and outline the physical limitations of the current study.

3.4.1. Effectiveness of Curriculum Learning and HER

Figure 4 presents the training success rate trajectories of all three experimental groups evaluated against the global task metric over 300,000 environment timesteps. The incremental contribution of each framework component is illustrated in Figure 7. Three mechanistically distinct phases are identifiable in the learning dynamics, each providing separable evidence for how the integration of progressive curriculum learning and HER resolves the cold-start problem and achieves superior sample efficiency relative to both baseline configurations.

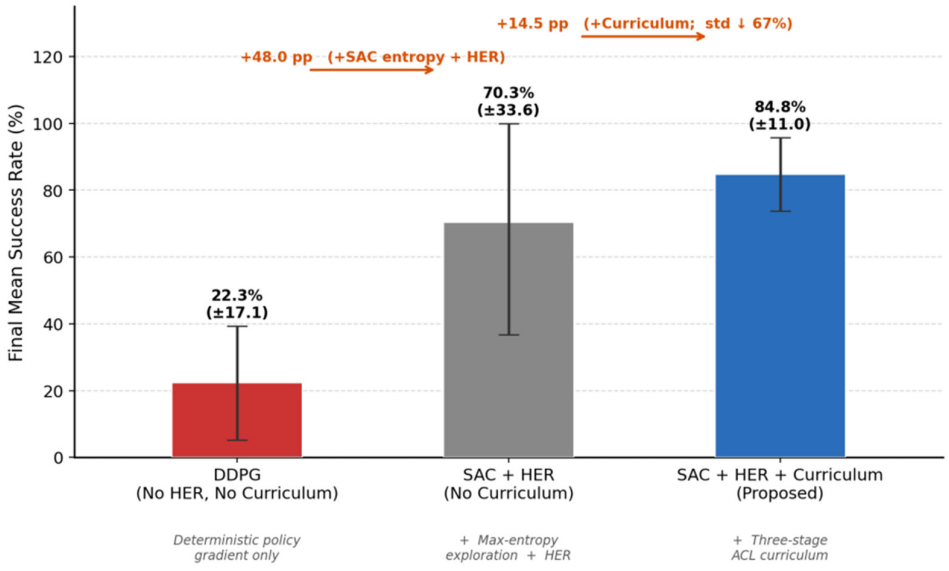


Figure 7. Ablation study showing incremental performance gain from each framework component. Error bars = ±1 std across 5 seeds. Arrows indicate incremental performance gains (in percentage points, pp) from each additional framework component.

Phase I—The Hidden Investment (0–100 k Steps). In the initial 100,000 timesteps, both Curriculum-SAC-HER and Standard SAC + HER register near-zero success rates on the global task metric—a superficial similarity that masks a fundamental divergence in the underlying learning process. For Standard SAC + HER, this plateau is consistent with genuine cold-start stagnation. Deployed directly into the full global workspace without staged task decomposition, the agent's entropy-regularized exploration generates trajectories that intersect the sparse goal region with negligible probability. Critically, this

absence of incidental successes starves the HER mechanism of the trajectories it requires to construct informative hindsight transitions. Without a positive reward signal to relabel, the replay buffer accumulates nearly exclusively uninformative negative transitions, the critic converges toward a near-uniform value landscape, and the actor cannot form coherent directional policy gradients. HER's sample efficiency advantage is therefore rendered latent—theoretically present but practically inaccessible in the absence of any goal-proximal experience. For Curriculum-SAC-HER, however, the global-task metric is a structurally misleading performance indicator during this window. The agent is not attempting the global task; it is mastering Stage 1 local sub-goals within a proximal bounding box of $R_1 = [0.02, 0.05, 0.02]^T$ meters. This physical restriction of the initial target distribution is the mechanism by which the curriculum directly resolves the cold-start problem: the narrow target volume substantially elevates the probability of the randomly exploring end-effector intersecting the goal region, reliably generating high-quality success trajectories from the earliest episodes of training. These trajectories immediately populate the HER replay buffer with informative hindsight transitions, activating HER's retrospective relabeling mechanism from the outset rather than leaving it idle. The actor network consequently begins accumulating a structured kinematic prior—a distributed representation of arm geometry, joint-space feasibility, and end-effector spatial correspondence—that would be entirely absent from a randomly initialized policy confronting the sparse global task directly. This constitutes the core sample efficiency advantage of Phase I: the curriculum bootstraps HER, and HER amplifies the informativeness of every collected transition, transforming an otherwise empty learning signal into a dense gradient stream.

Phase II—Breakout and Knowledge Transfer (~100–150 k Steps). The divergence between the two methods becomes unambiguous at approximately 100,000 timesteps, where the Curriculum-SAC-HER curve (solid blue) initiates a decisive breakout, ascending to approximately 38% success, while the Standard SAC + HER curve (dash–dot grey curve) remains near zero across its non-converging seeds. This asymmetry in emergence timing directly demonstrates that the kinematic prior accumulated during Stages 1 and 2 provides a substantive initialization advantage at the point of global-task exposure—an advantage that undirected exploration cannot replicate within the same timestep budget. The synergistic contribution of HER is particularly evident during the Stage 2 transition: as the target volume expands to $R_2 = [0.05, 0.15, 0.075]^T$, the agent's Stage 1 reaching trajectories systematically generate near-miss experiences at the boundaries of the new distribution. HER converts these near-misses into informative virtual successes, sustaining a continuous gradient signal across the expanded workspace without requiring additional completed goal-reaching episodes. This mechanism—curriculum-structured exploration supplying near-miss data and HER converting near-misses into constructive gradient signals—collectively accounts for a mean sample complexity reduction of 21.2% relative to Standard SAC + HER ($(260-205\text{ k})/260\ k \approx 21.2\%$, bootstrap 95% CI: 2.9–38.5%), where 205,000 steps represent the mean convergence timesteps across the four seeds satisfying the stability criterion.

At approximately 125,000 timesteps, corresponding to the transition into Stage 3 where the target bounding box expands to its maximum operational extent ($R_3 = [0.10, 0.30, 0.15]^T$) and the full global workspace is activated, the blue curve undergoes a transient regression to approximately 20%. This sudden expansion in spatial task difficulty—exposing the policy to goal positions across the full operational volume for the first time—is precisely what causes the temporary performance dip, as the policy must adapt its previously learned spatial representations to a strictly harder goal distribution. Rather than undermining the curriculum's value, however, this dip is precisely what a knowledge transfer hypothesis predicts: a bounded, momentary performance cost that tests the transferability of the

accumulated kinematic prior. The subsequent recovery—from 20% to approximately 85% within 50,000 steps—confirms that the agent adapts a pre-formed geometric prior rather than relearning task structure from scratch, and indicates that the kinematic representations accumulated during Stages 1 and 2 are sufficiently general to support rapid fine-tuning under the expanded distribution.

Phase III—Variance as the Definitive Discriminator (150–300 k Steps). In the final phase, Standard SAC + HER eventually produces a mean success rate of approximately 70.3%, which might superficially appear competitive with Curriculum-SAC-HER's 84.8%. The definitive discriminator, however, is not the mean but the variance. The confidence band surrounding the grey curve—a standard deviation of $\pm$33.6% spanning nearly the full success-rate range—is diagnostic of a policy acquisition process governed by initialization stochasticity rather than a reliable learning mechanism. Without structured sub-goals to constrain early exploration, whether a given seed converges is determined almost entirely by whether its initial weight configuration happens to generate trajectories that accidentally contact the sparse goal region and bootstrap HER. Seeds that miss this initialization-dependent convergence accumulate no meaningful gradient signal and converge to degenerate or catastrophically unstable policies. Curriculum-SAC-HER achieves a cross-seed standard deviation of $\pm$11.0% under identical evaluation conditions—a 67% compression of variance in point-estimate terms relative to Standard SAC + HER ($\pm$33.6% $\rightarrow$ $\pm$11.0%), though the small sample size (N = 5) precludes a statistically precise characterization of this variance ratio—demonstrating that the curriculum's structured progression substantially decouples final policy quality from initialization luck.

It should be noted that this residual $\pm$11.0% variance does not reflect catastrophic failure: no seed exhibits the uninformative value landscape under sparse rewards or irrecoverable policy collapse observed in the ablation group. Rather, it reflects moderate variation in Stage 3 convergence speed, arising from stochastic differences in the timing at which individual seeds satisfy the $\overline{S} \geq 0.8$ Stage 2 advancement criterion and thereby enter the global task distribution with differing remaining timestep budgets. At the extreme spatial boundaries of the Stage 3 global workspace, goal-proximal trajectory density is inherently lower, producing minor instability for seeds that transition to Stage 3 later in the training lifecycle. This distinction is consequential for interpreting the framework's reliability: the ACL module converts the catastrophic, irrecoverable failure variance of the ablation into moderate, bounded convergence-speed variance—a qualitatively superior outcome for practical deployment.

Taken together, the three phases establish a coherent and complete mechanistic account: curriculum learning resolves the cold-start problem by guaranteeing non-trivial HER signals from the first training episode; HER amplifies sample efficiency by extracting constructive gradient information from near-miss trajectories at each stage boundary; and the performance-driven stage transition criterion ensures that each spatial expansion occurs only after the prior distribution is fully consolidated, preventing catastrophic forgetting and compressing cross-seed variance to a moderate, recoverable regime.

3.4.2. Failure Analysis of Standard Baselines

The failure modes of the two baseline configurations are mechanistically distinct and, taken together, precisely delineate the theoretical gap that Curriculum-SAC-HER is designed to close. Analyzing these failures separately provides a rigorous account of why each component of the proposed framework—the curriculum scheduler, HER, and their integration—is a necessary rather than incidental design choice.

The Structural Failure of DDPG. The standard DDPG rapidly plateaus at approximately 22.3% mean success and exhibits no statistically meaningful upward trend for

the remainder of the 300,000-step budget. This outcome is structurally inevitable given the algorithm's exploration mechanism. DDPG's deterministic policy gradient relies on additive stochastic perturbations—typically Ornstein–Uhlenbeck process noise—to drive exploration. In a 7-DOF continuous joint space, the measure of trajectories that such temporally correlated, narrow-band noise can generate is severely constrained relative to the full configuration manifold. The probability of an Ornstein–Uhlenbeck-perturbed deterministic policy generating an end-effector trajectory that intersects the sparse goal region is, in practice, vanishingly small.

Consequently, the critic receives a near-uniformly negative reward signal across virtually all collected transitions. Without discriminative gradient information, the Bellman update cannot differentiate the value of goal-directed from goal-absent actions, causing the critic to converge toward a near-uniform value landscape. The actor, receiving a flat policy gradient, does not learn to freeze in order to minimize the step penalty—under the strictly sparse reward function ($r_t = -1$ for every non-success timestep), a stationary arm accumulates the maximum possible cumulative penalty and therefore cannot be interpreted as a cost-minimizing strategy. Rather, the observed oscillatory and near-static arm motions are symptomatic of an actor that receives no informative gradient signal from a collapsed critic, defaulting to low-magnitude, stochastic outputs governed by weight initialization noise rather than any acquired goal-directed competency. The residual 22.3% mean success rate most plausibly reflects the small fraction of episodes in which the robot's initial joint configuration places the end-effector within geometric proximity of the spawned goal, rather than any learned reaching skill—a conclusion consistent with the high per-seed variance (std: 17.1%) that would be expected if success is determined by episode initialization geometry rather than policy quality.

The initialization-dependent convergence of Standard SAC + HER: The failure of Standard SAC + HER is mechanistically distinct from, and theoretically more informative than, the DDPG collapse. SAC's maximum-entropy objective constitutes a theoretically superior exploration strategy, and the eventual mean success rate of 70.3% across surviving seeds confirms that entropy-regularized exploration is, in principle, sufficient for this task. The critical failure is not algorithmic inadequacy but structural fragility: the utility of both SAC's entropy bonus and HER's hindsight relabeling is entirely contingent on the agent's initial trajectories bearing some geometric proximity to the goal region.

This contingency exposes the cold-start problem in its most acute form. In the full global workspace of the 7-DOF task, with target positions uniformly distributed across a volume of approximately $0.2 \times 0.6 \times 0.3$ cubic meters, the probability that an undirected maximum-entropy policy generates end-effector trajectories that intersect the 0.05 m success radius is negligibly small. HER's retrospective relabeling mechanism, which depends on observed episode states to construct virtual goals, is consequently starved of the incidental near-success trajectories it requires to generate informative hindsight transitions. Without such transitions, the HER-augmented replay buffer remains populated almost exclusively with uniformly negative transitions, the critic value landscape converges toward a near-uniform uninformative state, and the actor cannot form coherent goal-directed gradients—the same collapse mechanism as DDPG, despite the fundamentally stronger exploration framework.

Whether a given seed escapes this cold-start trap is determined almost entirely by whether its initial weight configuration generates sufficient exploratory variance near the goal region to produce even a single accidental success within the first tens of thousands of timesteps. Seeds that satisfy this condition—Seeds 202, 204, and 205—bootstrap a positive learning cycle in which HER constructs increasingly informative transitions, the critic develops a meaningful value gradient, and the actor converges to high performance. Seeds

that do not—Seed 203 (final: 9.3%) and Seed 201 (final: 58.7%, following catastrophic forgetting at 230 k steps)—receive no such bootstrap signal and either stagnate permanently or collapse after transient convergence.

These two seeds represent mechanistically distinct failure modes. Seed 203 exhibits permanent stagnation from the outset: the initial weight configuration fails to generate any goal-proximal trajectories, HER receives no informative transitions to relabel, and the critic converges to a near-uniform uninformative value landscape from which recovery within the 300,000-step budget is essentially impossible. Seed 201, by contrast, achieves transient convergence to approximately 85% success before collapsing at approximately 230,000 timesteps—a qualitatively different trajectory consistent with catastrophic interference in the critic network. In the absence of curriculum structure, the Standard SAC + HER agent operating in the full global workspace encounters a non-stationary transition distribution throughout training: as the policy improves, the distribution of states near the goal shifts substantially, producing large gradient magnitudes in the critic update that can overwrite previously learned value estimates. For Seed 201, we hypothesize that this non-stationarity—compounded by the absence of stable kinematic priors from early training—produced a period of apparent convergence followed by destabilization of the critic's value estimates as new regions of the workspace were explored in later training. Once the critic's value landscape becomes corrupted in this manner, the actor loses coherent directional gradient signals and reverts toward uninformative behavior.

This mechanism is precisely what the replay buffer retention policy in Curriculum-SAC-HER structurally prevents: by maintaining early-stage kinematic transitions in the buffer across all stage boundaries, the critic continues to receive gradient signals from stable, low-difficulty transitions that function as an implicit value anchor, preventing the landscape corruption observed in Seed 201. This bifurcation mechanism generates the observed $\pm$33.6% standard deviation, which spans nearly the full success-rate range and is the quantitative signature of a learning process governed by initialization luck rather than algorithmic reliability. Future strategies to mitigate this initialization sensitivity without curriculum-based intervention include warm-start initialization using pre-trained sub-goal-reaching policies, population-based training across diverse initializations with selective pressure toward early successes, and learned difficulty estimators that dynamically adjust goal sampling based on online performance signals—each of which targets the cold-start vulnerability at its structural root.

Structural Resolution via Curriculum-SAC-HER: The Curriculum-SAC-HER framework addresses both failure modes at their theoretical root through a single architectural intervention: spatially constraining the initial target distribution. By restricting Stage 1 targets to a proximal bounding box of $R_1 = [0.02, 0.05, 0.02]^T$ meters, the curriculum scheduler eliminates the geometric lottery entirely. Every seed, regardless of weight initialization, is guaranteed to generate end-effector trajectories that intersect the goal region within the first episodes of training. HER immediately receives the high-quality success trajectories it requires to construct informative hindsight transitions, bootstrapping a stable positive learning cycle from the outset. This architectural guarantee transforms what is, under Standard SAC + HER, a stochastic initialization-dependent process into a deterministic stage-wise skill acquisition progression.

Equally importantly, the performance-driven stage transition criterion—advancement only upon $\overline{S} \geq 0.8$ over a 100-episode sliding window—ensures that the kinematic prior accumulated in Stage 1 is fully consolidated before the goal distribution expands. This prevents premature exposure to harder distributions that would otherwise reintroduce the cold-start vulnerability at each stage boundary. The empirical consequence is a compression of cross-seed variance from $\pm$33.6% under Standard SAC + HER to $\pm$11.0% under

Curriculum-SAC-HER, with one seed (Seed 405) failing to satisfy the three-consecutive-checkpoint stability criterion, and no seed exhibiting catastrophic collapse of the kind observed in the ablation group—demonstrating that structured geometric progression is the critical missing component in applying sparse-reward deep reinforcement learning to high-dimensional continuous manipulation.

3.4.3. Limitations and Future Work

Despite the demonstrated efficacy of Curriculum-SAC-HER within the simulated environment, two principal limitations define the current scope of this work and motivate concrete directions for future research. These limitations are not incidental but structurally inherent to the simulation design, and their honest characterization is essential for contextualizing the zero-shot robustness results reported in Section 3.3.

Limitation 1: Inertial Sensitivity and the Boundaries of the Learned Dynamic Prior: The zero-shot robustness evaluation in Section 3.3 reveals a pronounced and physically meaningful asymmetry in the policy's response to payload perturbations: performance degrades moderately but gracefully to 71.8% under a 1.5× mass increase, while a 0.5× mass reduction produces a disproportionately larger performance drop to 33.8%, characterized by end-effector overshoot rather than positional inaccuracy. This asymmetry warrants a mechanistic interpretation grounded in qualitative inertial dynamics reasoning.

The SAC maximum-entropy training objective, operating exclusively under the nominal inertial properties of the KUKA LBR iiwa arm throughout the curriculum, is hypothesized to converge toward a conservative, deceleration-biased torque policy: one that applies cautious approach profiles that, under nominal inertial conditions, reliably arrest end-effector motion within the success radius. This behavioral prior is a rational consequence of entropy maximization in a sparse-reward setting—the agent is incentivized to develop cautious, high-certainty approach trajectories that reliably satisfy the 0.05 m success threshold, rather than aggressive, high-velocity motions that risk overshooting the sparse goal region.

Under a 1.5× payload increase, the same torque commands—calibrated under nominal inertial conditions—produce reduced joint accelerations, slowing end-effector motion and increasing the number of timesteps required to reach the target. For the subset of trajectories where this slowdown remains within the 500-timestep episode budget, the conservative approach profile retains sufficient directional accuracy to complete the task, resulting in moderate but graceful degradation to 71.8%. Beyond the 1.5× threshold, the absolute torque ceiling of the manipulator prevents the policy from generating sufficient joint torques to overcome the increased gravitational loading, causing the success rate to plateau at approximately 13% as shown in Figure 6.

Under a 0.5× payload reduction, however, the same conservative torque profile is applied to a system with substantially lower inertia, generating excess kinetic energy relative to the learned deceleration model. The lighter end-effector overshoots the target zone, and the policy's learned stopping behavior—calibrated for a heavier load—cannot arrest the trajectory within the 0.05 m success radius. The resulting high variance (±32.5%) under light-load conditions reflects the sensitivity of overshoot magnitude to the specific trajectory geometry and approach angle across evaluation episodes. This behavioral asymmetry indicates that the current framework acquires a highly task-specific dynamic prior during curriculum training, rather than a load-agnostic kinematic policy.

Future work will address this limitation by incorporating online inertial parameter estimation as an auxiliary input to the policy network, enabling the agent to condition its deceleration profile on real-time estimates of end-effector inertia. Additionally, adaptive entropy scheduling—modulating the SAC temperature parameter α in response to

detected inertial mismatch—represents a promising direction for improving robustness under bidirectional payload perturbations without sacrificing nominal performance. A related concern is that the uniform -1 step penalty without smoothness regularization means that the learned policy is not verified to produce smooth torque trajectories; future work will incorporate torque-rate penalty terms of the form $\lambda\|a_t - a_{t-1}\|^2$ prior to any hardware deployment.

Limitation 2: Simulation Fidelity and the Sim-to-Real Gap: The entirety of the present study was conducted within the PyBullet physics engine, which, while providing a tractable and reproducible experimental platform, introduces several categories of abstraction that bound the validity of the parametric robustness evaluation conclusions. As comprehensively documented by Zhao et al. [18] in their survey of sim-to-real transfer methods—including domain randomization, domain adaptation, and meta-learning—the gap between simulated and real-world dynamics consistently degrades policy performance upon hardware deployment, and closing this gap remains an active frontier in DRL-based robotics research.

Most critically, PyBullet models joint torque control as an idealized, instantaneous actuation process operating at a simulation frequency of 240 Hz, with the high-level DRL policy executing at 60 Hz via an action repeat of 4. In contrast, real KUKA LBR iiwa hardware employs a 1 kHz joint-level servo control loop—a four-fold increase in temporal resolution relative to the simulated policy's effective control frequency. This discrepancy means that the simulated policy executes at a substantially coarser timescale than a physically deployed system, potentially masking high-frequency dynamic instabilities that would manifest only on real hardware.

Beyond control frequency, the simulation omits two categories of physical non-linearity that are known to substantially degrade policy transfer in practice. First, actuator latency—the 1–3 ms communication delay between the high-level policy and the joint-level servo—introduces a systematic lag between commanded and executed torques that the policy has no mechanism to anticipate or compensate for in the current formulation. Second, joint backlash—the mechanical compliance arising from gear train elasticity in the iiwa's harmonic drive actuators—introduces position hysteresis at direction reversals that is entirely absent from the simulated observation pipeline. Both effects produce stochastic perturbations in the effective state trajectory that the current proprioceptive state vector $\boldsymbol{s_t} \in \mathbb{R}^{20}$ cannot represent, creating a systematic mismatch between the state distributions encountered during simulation training and physical deployment. Furthermore, actuator latency and gear compliance in physical hardware introduce partial observability not captured by the current MDP formulation, motivating future POMDP-based or history-conditioned policy extensions in which the agent conditions its actions on a fixed-length history of proprioceptive observations rather than the instantaneous state vector alone.

Future work will address the Sim-to-Real gap through two complementary approaches. First, comprehensive Domain Randomization will be embedded directly within each curriculum stage during training, randomizing physical parameters including link masses, joint damping coefficients, and actuator gains and—critically—introducing time-correlated control noise models that simulate actuator latency and gear compliance across episodes. This approach aims to produce policies with sufficient distributional robustness for direct zero-shot deployment on physical 7-DOF hardware, bypassing the need for environment-specific fine-tuning. Second, the simulation environment will be upgraded to model actuator dynamics explicitly, including first-order lag models for motor response and backlash hysteresis for the harmonic drive joints, to reduce the fidelity gap at the source. To provide specificity regarding the proposed domain randomization strategy, the following physical parameters will be targeted in future work, selected based on documented sources

of simulation-to-hardware mismatch for the KUKA LBR iiwa platform. Link masses will be randomized within ±15% of nominal URDF values, expanding the current training range of ±10% to encompass manufacturing tolerances and unmodeled payload attachment variability. Joint damping coefficients will be randomized within [0.3×, 1.5×] of nominal values, reflecting viscous damping uncertainty arising from harmonic drive gear train lubrication variability in the iiwa's series-elastic actuators. Actuator gain perturbations will be randomized within ±20% of nominal values, modeling the controller parameter uncertainty inherent in hardware deployment without prior system identification. Control latency will be injected as a uniform random delay sampled from [0, 3] ms per timestep, simulating the communication delay between the high-level DRL policy and the iiwa's joint-level Fast Research Interface (FRI) servo loop operating at 1 kHz. Finally, zero-mean Gaussian observation noise will be added to joint angle and angular velocity measurements (σ = 0.01 rad and σ = 0.05 rad/s, respectively), reflecting the encoder resolution limits of the iiwa's optical joint sensors. These parameter ranges are informed by publicly available KUKA LBR iiwa technical specifications and prior sim-to-real transfer studies on comparable high-DOF platforms [18], and collectively target the principal sources of dynamic mismatch between the PyBullet simulation model and physical iiwa hardware.

Limitation 3: Obstacle-Free Workspace and Perception Constraints: The current state representation and reward formulation presuppose an unobstructed operational volume, constraining the framework's direct applicability to structured, free-space environments. The agent's proprioceptive-only state vector $\boldsymbol{s_t} \in \mathbb{R}^{20}$ provides no exteroceptive spatial awareness, rendering the policy incapable of reasoning about environmental geometry beyond the end-effector goal position. Practical manipulation tasks in industrial and assistive robotics settings routinely occur in cluttered workspaces where collision avoidance is a co-equal objective alongside goal reaching. Future curriculum designs will incorporate progressive intermediate stages populated with static and subsequently dynamic obstacles, building upon DRL frameworks for collision-avoidance trajectory planning under uncertain constraints [17], necessitating a transition from the current direct state extraction paradigm to the integration of exteroceptive spatial perception—specifically, depth image observations or point cloud encodings from simulated RGB-D sensors—alongside the existing proprioceptive state. Collision-penalty terms will be incorporated into the reward structure to provide explicit safety constraints, and the curriculum advancement criterion will be extended to require simultaneous satisfaction of both goal-reaching and collision-free trajectory objectives. This extended framework constitutes a principled roadmap toward Curriculum-SAC-HER deployment in realistic, cluttered manipulation scenarios involving compliant grasping and dynamic obstacle avoidance.

Limitation 4: Statistical Power and Seed Sample Size: The central empirical claim of this framework does not rest on mean performance comparisons, but on the structural compression of cross-seed variance: from ±33.6% under Standard SAC + HER—a range spanning nearly the full success-rate spectrum, diagnostic of a learning process governed by initialization stochasticity—to ±11.0% under Curriculum-SAC-HER, with no seed exhibiting catastrophic collapse. The reported p-values (p = 0.028, p = 0.008) are provided as supplementary indicators only. With N = 5 seeds per group, the Wilcoxon rank-sum test produces a highly discretized p-value distribution; these values should therefore not be interpreted as definitive proof of mean superiority. Future work will replicate findings across $N \geq 10$ seeds with bootstrap confidence intervals to establish statistically robust conclusions.

Limitation 5: Incomplete Ablation of Baseline Configurations. The DDPG baseline evaluated in this work does not incorporate HER, as the primary intent was to establish a lower bound demonstrating the structural inadequacy of deterministic exploration under

sparse rewards. A complete ablation including DDPG + HER would more precisely isolate the independent contribution of the SAC maximum-entropy objective from that of hindsight relabeling; this comparison is deferred to future work with an expanded computational budget.

4. Conclusions

This paper presents Curriculum-SAC-HER, a fusion framework integrating Soft Actor–Critic, Hindsight Experience Replay, and a performance-driven three-stage Automatic Curriculum Learning scheduler, for end-to-end continuous motion control of a 7-DOF redundant manipulator under strictly sparse rewards. Through progressive spatial task decomposition, the framework achieves a mean final success rate of 84.8% (std: 11.0%) across five independent random seeds, with four of five seeds exceeding 83%—converging at a mean of approximately 205,000 environment timesteps across the four seeds satisfying the stability criterion (bootstrap 95% CI: 160,000–252,500 steps), with one seed reaching the threshold only at the terminal checkpoint. This represents a mean reduction of approximately 21.2% in effective sample complexity relative to the Standard SAC + HER ablation (bootstrap 95% CI: 2.9–38.5%), with the confidence interval width reflecting the inherent variance of small-sample convergence estimation. The Standard DDPG baseline fails to exceed 22.3% within the full 300,000-step budget, confirming the structural inadequacy of deterministic exploration under sparse rewards in high-dimensional continuous spaces.

The central academic contribution of this work extends beyond the performance figures themselves. We identify the cold-start exploration bottleneck—the catastrophic dependence of sparse-reward convergence on stochastic weight initialization in high-dimensional action spaces—as the primary obstacle to reliable policy acquisition in this class of problems. Curriculum-SAC-HER resolves this bottleneck structurally: by constraining the initial target distribution to a proximal sub-space in Stage 1, the curriculum guarantees that every seed generates goal-proximal trajectories from the earliest training episodes, immediately bootstrapping HER with high-quality hindsight transitions and activating its sample efficiency advantage from the outset. The empirical consequence of this structural intervention is a compression of cross-seed performance variance from $\pm 33.6\%$ under Standard SAC + HER to $\pm 11.0\%$ under the proposed framework—a 67% reduction that reflects not merely higher average performance, but a qualitatively more reproducible and initialization-independent learning process. Critically, the residual $\pm 11.0\%$ variance does not represent catastrophic failure; no seed exhibits an uninformative value landscape under sparse rewards or irrecoverable policy collapse. It reflects moderate variation in Stage 3 convergence speed, a qualitatively superior failure mode for practical deployment.

Based on the empirical findings and framework analysis presented in this work, the ACL module is most beneficial in settings characterized by the following conditions: (1) high-dimensional continuous action spaces (e.g., $\geq$6 DOF) where undirected exploration has negligible probability of generating goal-proximal trajectories, as quantified by the volumetric analysis in Section 1; (2) strictly sparse reward signals that provide no gradient information until a binary success threshold is crossed, rendering standard Bellman updates uninformative in early training; (3) tasks admitting a natural spatial decomposition, i.e., where the global workspace can be partitioned into nested sub-regions of increasing difficulty without requiring manual reward engineering; and (4) reproducibility-critical applications where cross-seed training variance must be structurally bounded rather than managed post hoc through seed selection or result filtering. Conversely, the ACL design overhead may not be justified for tasks with dense reward shaping, low-dimensional action spaces, or goal distributions that do not permit meaningful geometric partitioning.

Future research will proceed along two corresponding directions. To bridge the Sim-to-Real gap, subsequent work will embed comprehensive Domain Randomization directly within each curriculum stage, randomizing physical parameters including link masses, joint damping coefficients, and actuator gains, and incorporating time-correlated control noise models that simulate actuator latency and gear compliance across training episodes. This approach targets sufficient distributional robustness as a step toward, but not yet achieving, zero-shot evaluation on physical 7-DOF hardware without environment-specific fine-tuning. To address workspace complexity, the curriculum architecture will be extended with obstacle-populated intermediate stages, integrating exteroceptive spatial perception—specifically RGB-D sensor inputs and point cloud encodings—alongside explicit collision-penalty reward terms, to enable safe, generalizable manipulation in cluttered environments involving dynamic obstacle avoidance and compliant grasping. Together, these extensions constitute a principled roadmap toward robust, real-world deployment of curriculum-structured deep reinforcement learning for dexterous manipulation.

Author Contributions: Conceptualization, Y.Z.; Methodology, Y.Z.; Software, Y.Z.; Investigation, Y.Z.; Formal analysis, Y.Z.; Writing—original draft, Y.Z.; Writing—review and editing, Y.Z. and J.G.; Supervision, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and code supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors used Claude (Anthropic, https://claude.ai, accessed on 20 June 2026) and Google Gemini (Google, https://gemini.google.com, accessed on 20 June 2026) to assist with language polishing and manuscript revision.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

DRL	Deep Reinforcement Learning
SAC	Soft Actor–Critic
HER	Hindsight Experience Replay
ACL	Automatic Curriculum Learning
DOF	Degree of Freedom
MLP	Multi-Layer Perceptron
DDPG	Deep Deterministic Policy Gradient
CHER	Curriculum-guided Hindsight Experience Replay
HGG	Hindsight Goal Generation
URDF	Unified Robot Description Format
API	Application Programming Interface

References

1. Kroemer, O.; Niekum, S.; Konidaris, G. A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms. *J. Mach. Learn. Res.* **2021**, *22*, 1–82.
2. Ibarz, J.; Tan, J.; Finn, C.; Kalakrishnan, M.; Pastor, P.; Levine, S. How to Train Your Robot with Deep Reinforcement Learning: Lessons We Have Learned. *Int. J. Robot. Res.* **2021**, *40*, 698–721. [CrossRef]
3. 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, Stockholm, Sweden, 10–15 July 2018.
4. Tang, C.; Abbatematteo, B.; Hu, J.; Chandra, R.; Martín-Martín, R.; Stone, P. Deep Reinforcement Learning for Robotics: A Survey of Real-World Successes. *Annu. Rev. Control Robot. Auton. Syst.* **2025**, *8*, 153–188. [CrossRef]
5. Zhang, T.; Mo, H. Towards Multi-Objective Object Push-Grasp Policy Based on Maximum Entropy Deep Reinforcement Learning under Sparse Rewards. *Entropy* **2024**, *26*, 416. [CrossRef] [PubMed]

6. Fang, M.; Zhou, T.; Du, Y.; Han, L.; Zhang, Z. Curriculum-Guided Hindsight Experience Replay. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019*; Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R., Eds.; Neural Information Processing Systems Foundation Inc.: San Diego, CA, USA, 2019; pp. 12602–12613.
7. Ren, Z.; Dong, K.; Zhou, Y.; Liu, Q.; Peng, J. Exploration via Hindsight Goal Generation. In *Advances in Neural Information Processing Systems, Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, NeurIPS 2019*; Neural Information Processing Systems Foundation Inc.: San Diego, CA, USA, 2019; Volume 32.
8. Bing, Z.; Brucker, M.; Morin, F.O.; Li, R.; Su, X.; Huang, K.; Knoll, A. Complex Robotic Manipulation via Graph-Based Hindsight Goal Generation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 7863–7876. [CrossRef] [PubMed]
9. Sayar, E.; Iacca, G.; Knoll, A. Curriculum Learning for Robot Manipulation Tasks With Sparse Reward Through Environment Shifts. *IEEE Access* **2024**, *12*, 46626–46635. [CrossRef]
10. Bing, Z.; Zhou, H.; Li, R.; Su, X.; Morin, F.O.; Huang, K.; Knoll, A. Solving Robotic Manipulation With Sparse Reward Reinforcement Learning Via Graph-Based Diversity and Proximity. *IEEE Trans. Ind. Electron.* **2023**, *70*, 2759–2769. [CrossRef]
11. Paolo, G.; Coninx, M.; Laflaquière, A.; Doncieux, S. Discovering and Exploiting Sparse Rewards in a Learned Behavior Space. *Evol. Comput.* **2024**, *32*, 275–305. [CrossRef] [PubMed]
12. Or, K.; Wu, K.; Nakano, K.; Ikeda, M.; Ando, M.; Kuniyoshi, Y.; Niiyama, R. Curriculum-Reinforcement Learning on Simulation Platform of Tendon-Driven High-Degree of Freedom Underactuated Manipulator. *Front. Robot. AI* **2023**, *10*, 1066518. [CrossRef] [PubMed]
13. Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Abbeel, P.; Zaremba, W. Hindsight Experience Replay. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
14. Plappert, M.; Andrychowicz, M.; Ray, A.; McGrew, B.; Baker, B.; Powell, G.; Schneider, J.; Tobin, J.; Chociej, M.; Welinder, P.; et al. Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research. *arXiv* **2018**, arXiv:1802.09464v2.
15. 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.
16. Prianto, E.; Kim, M.; Park, J.-H.; Bae, J.-H.; Kim, J.-S. Path Planning for Multi-Arm Manipulators Using Deep Reinforcement Learning: Soft Actor–Critic with Hindsight Experience Replay. *Sensors* **2020**, *20*, 5911. [CrossRef] [PubMed]
17. Chen, L.; Jiang, Z.; Cheng, L.; Knoll, A.C.; Zhou, M. Deep Reinforcement Learning Based Trajectory Planning Under Uncertain Constraints. *Front. Neurorobot.* **2022**, *16*, 883562. [CrossRef] [PubMed]
18. Zhao, W.; Queralta, J.P.; 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)*; IEEE: Canberra, ACT, Australia, 2020; pp. 737–744.