Learning In-Hand Object Reaching to General 6D Poses

Junxiao Lin[2,1,3], Tianyue Wu[2,3], Jie Yin[2], Jia Pan[3], Kaifeng Zhang[2], Weiming Zhi[1,*]

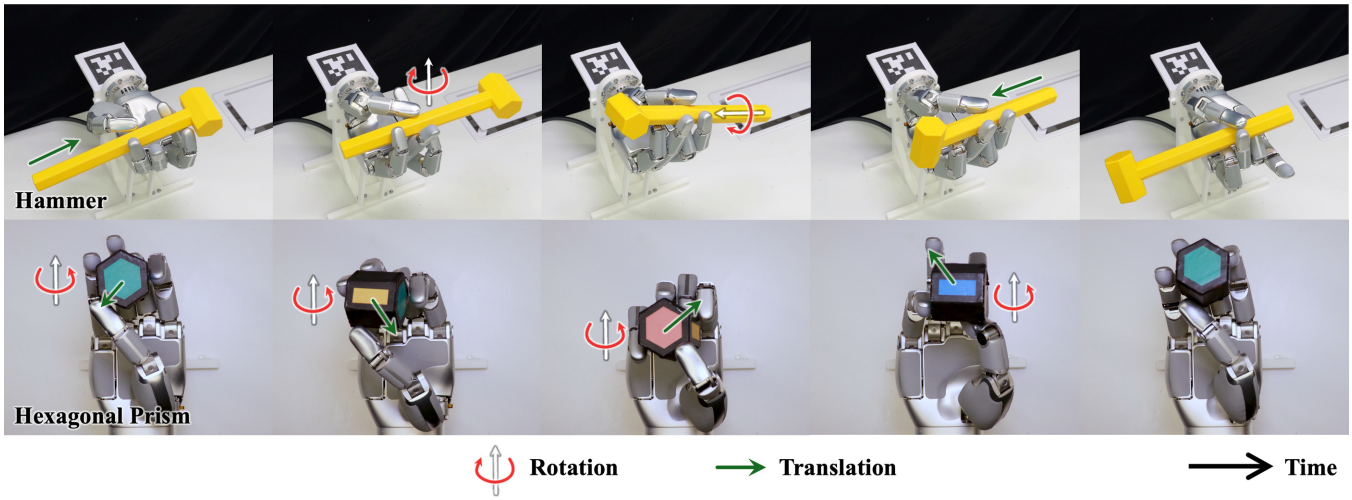

Fig. 1. Real-world in-hand 6D pose reaching. The Hammer (top) and Hexagonal Prism (bottom) are translated and rotated within the hand toward commanded poses. Green and red arrows indicate translational and rotational motions, respectively.

***Abstract*— In-hand manipulation allows multi-fingered dexterous hands to reconfigure grasped objects without releasing and regrasping them. This improves manipulation efficiency by reducing repeated grasp acquisition and large arm motions. However, most learning-based methods focus on reorientation, continuous rotation, or translation, whereas many tasks require joint control of object position and orientation. We formulate this capability as in-hand 6D object pose reaching: starting from an existing grasp, coordinated finger motions move the object to a palm-relative target pose. We present POISE (Palm-relative Object reaching In $\mathrm{SE}(3)$), a sim-to-real reinforcement learning framework for this task. POISE combines diverse stable-grasp initialization, goal- and geometry-conditioned control, an adaptive 6D goal curriculum, and a compact reward scheme for pose reaching and grasp preservation. In simulation, diverse initialization raises held-out-grasp success from 40.1% to 51.5% and post-drop recovery from 33.8% to 72.9%; the curriculum raises full-range success from 6.2% to 59.5%. On hardware, the grasp-maintenance reward improves three-target sequence success from 20% to 80%. In real-world experiments, POISE reaches successive 6D targets without manual reset across multiple object geometries and wrist orientations, and recovers from external disturbances. To support further research in dexterous manipulation, we will release our code at `https://junxiaolin.github.io/poise-website/`.**

I. Introduction

Many manipulation tasks begin, rather than end, once an object has been grasped. While parallel-jaw grippers are effective for establishing stable grasps, they have limited ability to reconfigure an object after grasp acquisition. High-degree-of-freedom dexterous hands, in contrast, can continue to manipulate a grasped object through coordinated multi-finger motions. This in-hand dexterity is crucial when a robot needs to shift the grasping region on a tool, expose a desired contact surface, or align a functional axis with the environment.

We formulate this capability as in-hand 6D object pose reaching: given a grasped object and a random target pose defined relative to the hand in the special Euclidean group $\mathrm{SE}(3)$, a single multi-fingered dexterous hand must bring the object to the target using only in-hand manipulation. Unlike in-hand reorientation or translation-only tasks [1, 2], this formulation jointly constrains the object's relative 3D translation and 3D rotation. This coupled objective significantly constrains the feasible motion space and often requires contact switching or finger gaiting while maintaining a stable grasp. Successive reaching further requires each attained object pose to be supported by a grasp that permits continued manipulation.

Prior work has addressed aspects of this problem from two main directions. Model-based methods can achieve accurate local in-hand pose control, but they often rely on predefined contact modes, fixed contact sequences, or known and simplified object geometry [3]. As a result, they do not readily scale to random $\mathrm{SE}(3)$ reaching, where the hand must actively switch contacts and adapt to diverse object shapes. Reinforcement learning (RL), on the other hand, has become a powerful tool for dexterous in-hand manipulation because it can discover contact-rich behaviors without requiring high-quality demonstrations, which are difficult to acquire for such tasks. However, most learning-

[1] School of Computer Science, The University of Sydney, Australia. [2] Sharpa. [3] School of Computing and Data Science, The University of Hong Kong, HKSAR. * Corresponding author.

based in-hand manipulation work focuses on reorientation, continuous rotation, or constrained translation [1, 2, 4, 5]. Prior full-pose in-hand reaching methods remain confined to object-specific simulation [6, 7].

To bridge these gaps, we develop **POISE** (**P**alm-relative **O**bject reaching **I**n $\mathbf{SE}(3)$), a sim-to-real RL framework based on Proximal Policy Optimization (PPO) [8]. POISE comprises five components: diverse physics-validated stable-grasp resets for broad state coverage; a goal- and geometry-conditioned policy for shape-dependent control; an adaptive 6D goal curriculum for learning large pose changes; a compact reward scheme that combines pose reaching, goal completion, grasp maintenance, and regularization; and domain randomization for robust sim-to-real transfer. Experiments demonstrate generalization across initial grasps, recovery after the original grasp is lost, and continued 6D reaching across multiple object geometries and wrist orientations.

The main contributions of this letter are as follows:

1) A diverse, physics-validated stable-grasp initialization that improves generalization to unseen grasps and recovery after the original grasp is lost.
2) An RL recipe for in-hand 6D pose reaching that combines geometry-conditioned control and adaptive goal curricula with a grasp-maintenance reward based on balanced wrench coverage.
3) Real-world validation of successive 6D reaching across object geometries and wrist orientations, including disturbance recovery and a quantitative evaluation of the grasp-maintenance reward.

II. Related Work

Dexterous in-hand manipulation has been extensively studied, from early analyses of rolling and finger gaiting [9] to modern model-based and learning-based approaches. Model-based methods compute manipulation motions from explicit models of hand kinematics, fingertip contacts, and object motion. Motion-cone planning derives feasible object motions from contact constraints [10], while trajectory optimization and multi-modal planning coordinate in-grasp motion, finger gaiting, and compliant contact transitions [3, 11]. Model predictive control instead plans over locally identified or contact-implicit dynamics [12]–[14]. These approaches can enforce physical and kinematic constraints directly, but often depend on accurate models, predefined contact modes, or time-consuming online optimization. Extending them to a broad range of 6D goals and object geometries remains challenging.

Learning-based methods instead discover contact-rich multi-finger coordination through trial-and-error interaction [15]–[18]. RL has enabled real-world reorientation, continuous rotation, and constrained translation [1, 2, 4, 19]–[27]. Recent work further improves generalization across object geometry through visual or explicit shape conditioning [5, 28, 29]. Nevertheless, most of these methods control orientation or a constrained translational motion rather than a general 6D target pose.

Closest to our setting, Plappert et al. [6] introduced goal-pose-conditioned RL simulation environments for in-hand manipulation of block, egg, and pen objects. Initial solutions to these challenging tasks achieved only low success rates; Charlesworth and Montana [7] later improved performance using offline trajectory optimization, but still only in simulation. More recent systems learn 6D object pose reaching through joint arm–hand control toward world-frame goals [30]–[32]. Other controllers track human or synthetic hand–object reference motions [33]–[36]. In contrast, POISE reaches palm-relative $\mathrm{SE}(3)$ object goals through finger motions alone, without hand-motion references.

III. Problem Formulation

We consider a fixed-wrist hand moving an already-grasped object to commanded poses using only finger motions. Let $\mathcal{H}$ and $\mathcal{O}$ denote the palm- and object-fixed frames. The current and target palm-relative poses are ${}^{\mathcal{H}}\mathbf{T}_{\mathcal{O},t} = ({}^{\mathcal{H}}\mathbf{p}_t, {}^{\mathcal{H}}\mathbf{R}_t)$ and ${}^{\mathcal{H}}\mathbf{T}_{\mathcal{O},g} = ({}^{\mathcal{H}}\mathbf{p}_g, {}^{\mathcal{H}}\mathbf{R}_g) \in \mathrm{SE}(3)$. Their position and orientation errors are

$$\begin{aligned} e_p &= \left\| {}^{\mathcal{H}}\mathbf{p}_t - {}^{\mathcal{H}}\mathbf{p}_g \right\|_2, \\ e_R &= d_{\mathrm{SO}(3)}\left({}^{\mathcal{H}}\mathbf{R}_t, {}^{\mathcal{H}}\mathbf{R}_g\right) \\ &= \frac{1}{\sqrt{2}} \left\| \mathrm{Log}\left({}^{\mathcal{H}}\mathbf{R}_g ({}^{\mathcal{H}}\mathbf{R}_t)^\top\right) \right\|_{\mathrm{F}}. \end{aligned} \tag{1}$$

Here, $\mathrm{Log}(\cdot)$ denotes the matrix logarithm and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm; thus, $e_R \in [0, \pi]$ is the geodesic rotation angle between the two orientations.

We formulate the task as a goal-conditioned partially observable Markov decision process (POMDP). An MLP actor receives three-frame histories of the current and commanded finger-joint positions, estimated palm-frame object pose and velocity, the gravity direction expressed in the palm frame, a confidence score for the visual pose estimate, the target pose, the current-to-target translation and rotation, and an object-geometry descriptor $\mathbf{z}_{\mathcal{O}}$ (Sec. IV-B). The critic additionally receives noise-free simulator states, fingertip states, applied joint torques, and contact forces. The normalized action incrementally updates the commanded finger joints,

$$\mathbf{q}_{t+1}^{\mathrm{cmd}} = \mathrm{clip}\left(\mathbf{q}_t^{\mathrm{cmd}} + s_a \mathbf{a}_t, \mathbf{q}_{\min}, \mathbf{q}_{\max}\right). \tag{2}$$

A target is reached when both errors remain within their tolerances for a prescribed number of control steps. A new target is then sampled without resetting the hand–object state during training, requiring successive 6D reaches within one continuous rollout.

IV. Method

In-hand 6D pose reaching requires the policy to handle varied grasps and object geometries, explore a broad range of 6D poses, and preserve a stable grasp after reaching. POISE addresses these challenges through diverse stable-grasp initialization, geometry-conditioned control, an adaptive 6D goal curriculum, a grasp-preserving reward, and domain randomization, as summarized in Fig. 2.

A. Diverse Stable-Grasp Initialization

Because the task starts after grasp acquisition, each episode must begin from a stable hand–object state. With

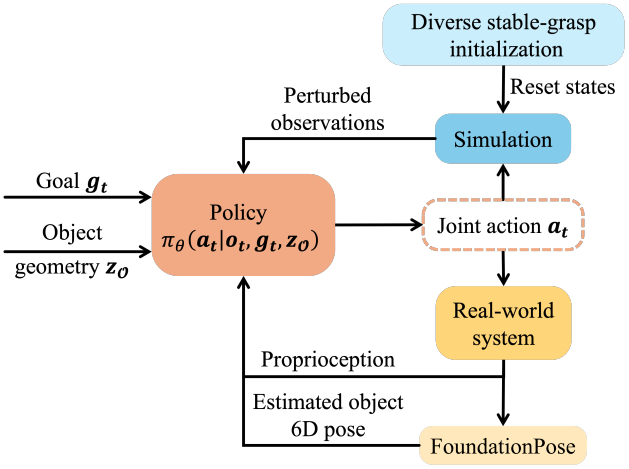

Fig. 2. Overview of POISE. Diverse stable grasps initialize simulation training, while FoundationPose supplies object-pose feedback during real-world deployment.

only one reset grasp, rollout exploration is largely restricted to states that the current policy can reach from that configuration, leaving alternative object–palm poses and contact arrangements underrepresented. We therefore initialize training from a diverse set of stable grasps.

Following the optimization-based grasp synthesis paradigm [37], we generate these grasps using a procedure adapted from [38]. Object poses are first sampled over a feasible in-hand workspace. For each pose, we optimize stable contacts and solve for the corresponding joint configuration $\mathbf{q}$; the optimized contact forces are then converted into an executable position-control command $\mathbf{q}^{\mathrm{cmd}}$. Randomized contact search produces multiple postures for each object pose. We then perturb the object pose, $\mathbf{q}$, and $\mathbf{q}^{\mathrm{cmd}}$, and simulate each candidate for 2 s under gravity and position control. Only candidates that retain the object are stored in the reset cache $\mathcal{G}$.

Sampling reset states from $\mathcal{G}$ makes grasp diversity an explicit part of the training distribution, rather than leaving it to be discovered through rollout exploration. We evaluate its effect on held-out grasps and recovery in Sec. V-B.1.

B. Geometry-Conditioned Policy

Object shape determines which surfaces and edges can support contact and how finger motion is transferred to the object. Thus, identical current and target poses can require different finger coordination for different geometries. We expose this information through an explicit shape descriptor.

We use the basis point set (BPS) representation [29, 39]. A fixed set of query points is shared by all objects in their canonical frames, and each entry records the normalized distance from one query point to the nearest object surface. The resulting ordered vector $\mathbf{z}_{\mathcal{O}}$ has the same dimension and spatial structure across objects.

The actor receives $\mathbf{z}_{\mathcal{O}}$ together with its observation and goal. During multi-object training, one actor is optimized jointly across objects without a categorical identity input; it must therefore use geometry to adapt its finger coordination. The BPS dimension is given in Sec. V-A.

C. Adaptive 6D Goal Curriculum

Sampling the full range of 6D goals from the outset makes early exploration difficult. We therefore use separate curricula that expand the maximum rotation and translation magnitudes from 5° and 10 mm to 180° and 30 mm, respectively. Only the magnitudes are scheduled; axes and directions remain randomized, with the same limits applied to all targets in a rollout. Targets outside the palm-frame workspace or intersecting the fixed hand base are resampled.

At each stage, magnitudes are sampled near the frontier, from the exposed range, or at the current limit in a $0.6/0.3/0.1$ ratio. When frontier success reaches 0.4, the rotation or translation limit increases by 10° or 10 mm. Each object maintains independent curriculum progress.

D. Grasp-Preserving Reward Design

Reaching the target pose does not guarantee a useful terminal grasp: the object may be loosely supported or held by too few contacts to continue manipulation. Our reward therefore combines pose reaching, verified goal completion, grasp maintenance, and actuation regularization.

Pose reaching. Using the errors in Eq. (1), the dense pose reward is

$$r_t^R = \exp\Big(-\frac{e_{R,t}}{30^\circ}\Big), \tag{3}$$

$$r_t^p = \alpha(e_{R,t})\,\exp\Big(-\frac{e_{p,t}}{\sigma(e_{R,t})}\Big), \tag{4}$$

$$r_t^{\mathrm{pose}} = \tfrac{1}{6} r_t^R + \tfrac{1}{4} r_t^p, \tag{5}$$

where α and σ vary with e_R as

$$(\alpha, \sigma) = \begin{cases} (0.15,\ 30\ \mathrm{mm}), & e_R > 45^\circ, \\ (0.40,\ 20\ \mathrm{mm}), & 25^\circ < e_R \le 45^\circ, \\ (1.00,\ 15\ \mathrm{mm}), & e_R \le 25^\circ. \end{cases} \tag{6}$$

When the orientation error is large, the position term is broad and downweighted, allowing translation needed for contact rearrangement. As the object becomes rotationally aligned, the reward increasingly emphasizes accurate positioning.

Goal completion. Let $s_t \in \{0, 1\}$ indicate that the object is retained and lies within both pose tolerances. A goal is verified only after this condition holds for $N = 20$ consecutive steps; $c_t \in \{0, 1\}$ marks the first step on which verification occurs. We use

$$r_t^{\mathrm{goal}} = 0.5 s_t + 45 c_t. \tag{7}$$

The per-step term encourages the policy to remain inside the target region, while the one-time bonus rewards verified completion.

Grasp maintenance. We approximate each active contact by four friction-cone wrench rays, with moments normalized by object size. For each force and torque axis, $m_{t,k} \in [0, 1]$ is the smaller of the maximum ray projections in its positive and negative directions. We aggregate the six margins as

$$Q_t = \left[\frac{1}{6}\sum_{k=1}^{6}(m_{t,k} + \varepsilon)^{-8}\right]^{-1/8} - \varepsilon. \tag{8}$$

This generalized mean emphasizes the weakest of the six bidirectional projection margins, encouraging balanced wrench coverage along the three force and three torque

axes. Contact contributions saturate with force, preventing the policy from increasing the score merely by squeezing harder.

Since attainable quality depends on the object and initial grasp, the reward preserves quality relative to the episode's stable reset rather than using one absolute threshold. Let $Q^\star$ be the mean quality over the first four control steps, clipped to $[0.08, 0.35]$. We define

$$r_t^{\text{grasp}} = 1 - \text{clip}\left(\frac{(Q^\star - Q_t)_+}{Q^\star}, 0, 1\right)^2. \tag{9}$$

The reward remains maximal while the initial wrench coverage is retained and decreases smoothly as the grasp becomes less secure.

Drop and actuation regularization. Let $d_t \in \{0, 1\}$ indicate a drop. To discourage object loss and unnecessarily aggressive control, the complete reward penalizes drops, joint torque, and instantaneous mechanical power:

$$\begin{aligned} r_t = r_t^{\text{pose}} + r_t^{\text{goal}} + 0.05 r_t^{\text{grasp}} \\ - 80 d_t - 0.10 \|\boldsymbol{\tau}_t\|_2^2 - 0.15 \left(\boldsymbol{\tau}_t^\top \dot{\mathbf{q}}_t\right)^2. \end{aligned} \tag{10}$$

E. Domain Randomization

Real deployment introduces both dynamics mismatch and imperfect visual pose feedback. We randomize object properties, friction, controller gains, and external disturbances to improve robustness to the former. For the latter, the actor receives joint and object-pose noise, pose delay, and hold-last dropouts; object velocities are estimated from the corrupted pose history. Table I summarizes the ranges.

TABLE I

SIM-TO-REAL RANDOMIZATION USED DURING TRAINING. $\mathcal{U}$ AND LOGU DENOTE UNIFORM AND LOG-UNIFORM SAMPLING.

Quantity	Range or distribution
Object scale	$\mathcal{U}[0.975, 1.025]$
Object mass	10–100 g
Inertia multiplier	$\text{LogU}[1, 4]$
CoM offset	Up to 5 mm per axis
Friction multiplier	$\mathcal{U}[0.5, 2.0]$
PD-gain multipliers	$\mathcal{U}[0.5, 2.0]$
Wrist orientation	Random over the full orientation range
External disturbances	$\sigma_a = 6.67$ m/s^2; $\sigma_\alpha = 133.33$ rad/s^2
Disturbance probability	$\text{LogU}[1/3000, 1/30]$ per step
Joint-position noise	$\mathcal{U}[-0.02, 0.02]$ rad
Object-pose noise	$\sigma_p = 10$ mm/axis; $\sigma_R = 5^\circ$ (clip 15°)
Pose delay	0–3 steps (0–150 ms)
Pose dropout	3% per step; hold last pose

V. EXPERIMENTS

We use simulation and hardware experiments for complementary purposes. Simulation provides controlled, quantitative evidence for the three main design choices: diverse stable-grasp initialization, the adaptive goal curriculum, and the grasp-maintenance reward. Hardware experiments instead focus on system-level capabilities, including reaching across object geometries and wrist orientations, recovery from external disturbances, and continued operation after reaching.

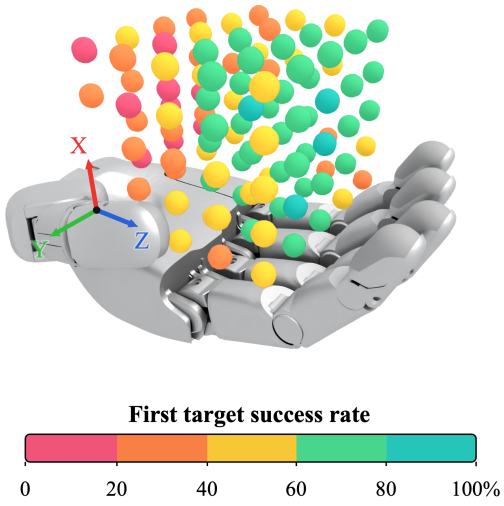

Fig. 3. First-target success of the diverse policy across independently generated held-out initial grasps under paired 30-mm/180° targets. Sphere positions indicate the initial object positions in the palm frame, and color indicates success rate.

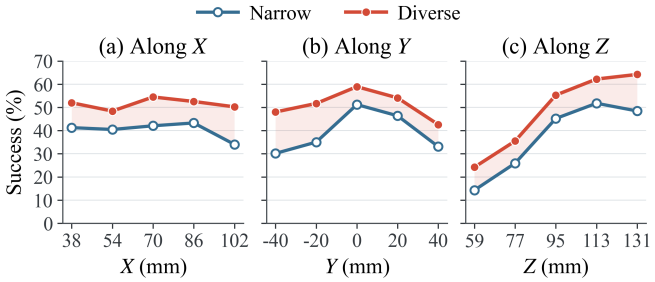

Fig. 4. First-target success of the narrow and diverse policies, marginalized along each palm-frame position axis. Palm-frame X is approximately normal to the palm, while Y and Z lie in the palm plane.

A. Experimental Setup

Simulation. We train the policies with asymmetric PPO in Isaac Lab [40]. Physics is simulated at 120 Hz, and the policy produces incremental joint-position commands at 20 Hz. Both actor and critic are $[512, 256, 128]$ ELU MLPs, and object geometry is represented by a 64-dimensional BPS descriptor. Unless noted otherwise, ablations use a 40-mm Hexagonal Prism and fixed palm-up wrist, changing only the component under study. A target is considered reached when the position and orientation errors remain below 10 mm and 10°, respectively, for 20 consecutive policy steps. After reaching, a new target is issued without resetting the hand–object state. Each episode lasts 400 policy steps and terminates early if the object is dropped.

Hardware. The real system consists of a 22-DoF Sharpa Wave hand [41] and a RealSense D435 camera [42]. FoundationPose [43] estimates the object pose from the RGB-D stream and a known object mesh. A calibrated camera-to-hand transform expresses the estimate in the palm frame, and the policy runs in closed loop at 20 Hz. The same observation and action interfaces are used in simulation and on hardware; only deployable proprioceptive and visual estimates are provided to the actor. The domain randomization used for transfer is summarized in Table I.

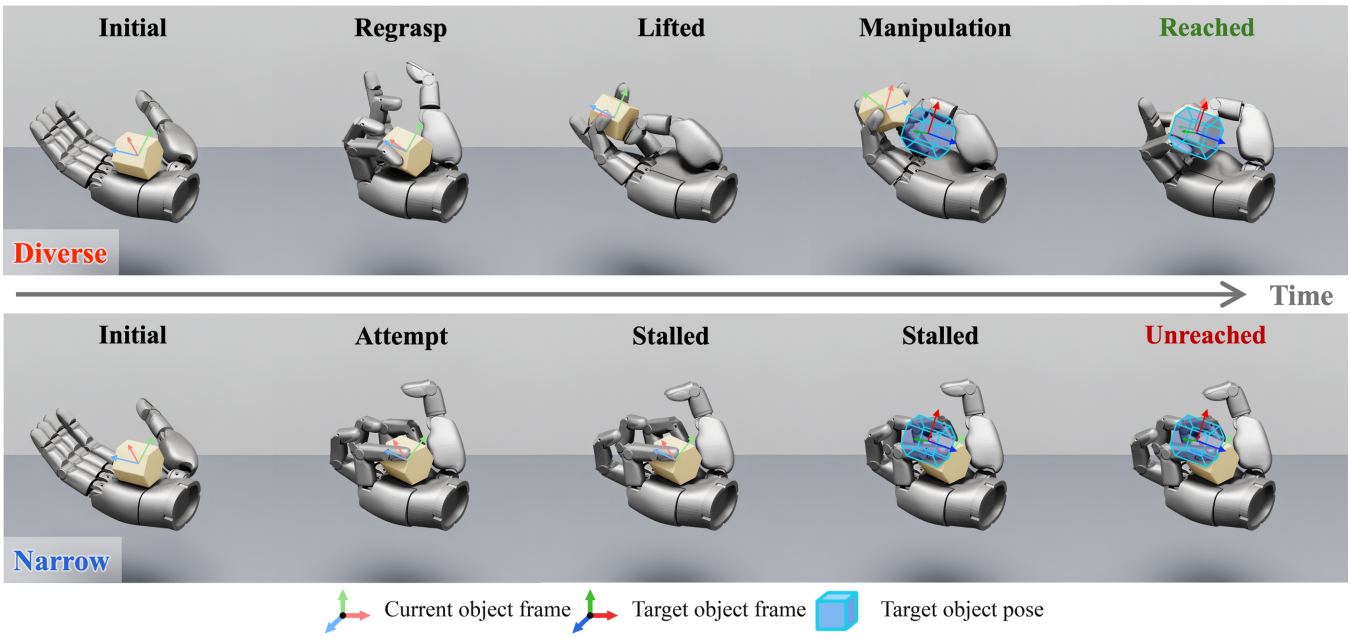

Fig. 5. Simulation recovery from the same post-drop state and target. The diverse policy (top) rebuilds contact, lifts the object from the palm, and reaches the target. The narrow policy (bottom) attempts to regrasp the object but stalls before reaching the target.

B. Quantitative Simulation Studies

1) Generalization Across Initial Grasps: To isolate the effect of diversity in grasp initialization, we compare the proposed method with a narrow variant whose reset cache is obtained by perturbing the joint state and object pose of a single grasp seed. Each cache contains 6,842 stable reset states; the conditions differ only in their reset-state distributions. We test whether this coverage supports reaching from unseen stable grasps and recovery after the original grasp is lost.

On independently generated held-out stable grasps, both policies receive the same initial states and 30-mm/180° first targets. Diverse initialization increases first-target success from 40.1% to 51.5% and the mean number of completed goals per episode from 0.69 to 1.00. Figs. 3 and 4 further characterize this gain over the initial object position. Smaller Z coordinates and the boundaries of the Y range are more difficult for both policies, while success varies less along X. The diverse policy has higher observed success in every marginalized position bin shown, suggesting that the gain is distributed across the held-out grasp workspace.

The recovery test then examines whether this broader state coverage also improves recovery after a grasp is broken. Both policies are evaluated on the same 720 trials, with the object resting on the palm and no established grasp. For 30-mm/180° targets that require lifting the object away from the palm, diverse initialization increases first-target success from 33.8% to 72.9%. Among successful trials, the mean time to recover and reach the target decreases from 11.63 s to 9.45 s. Figure 5 contrasts representative rollouts from the same post-drop state and target: the diverse policy rebuilds multi-finger contact, lifts the object, and reaches the target, whereas the narrow policy stalls without reaching the target. Together, the two tests indicate that broader coverage during training improves both generalization across grasp poses and recovery after a drop.

TABLE II

ABLATION OF THE ADAPTIVE GOAL CURRICULUM ON THE HEXAGONAL PRISM. STRICT AND NEAR DENOTE FIRST-TARGET SUCCESS UNDER THE 10°/10-MM AND 20°/20-MM CRITERIA, RESPECTIVELY.

	Full range			Maximum change		
Training	Strict	Near	Goals/ep.	Strict	Near	Goals/ep.
Without curriculum	6.2%	18.8%	0.08	3.4%	13.9%	0.04
Curriculum	**59.5%**	**77.0%**	**1.55**	**55.3%**	**72.8%**	**1.08**

2) Effect of the Adaptive Goal Curriculum: We compare our curriculum with a variant that samples target magnitudes uniformly over the full range throughout training. Each training run uses 64,000 parallel simulation environments. We evaluate the policies from the same stable grasps and with the same first targets in two suites: full-range targets with rotations in $[5°, 180°]$ and translations in $[10, 30]$ mm, and maximum-change targets of 180° and 30 mm.

We report strict first-target success under the standard 10°/10-mm criterion and near success under a relaxed 20°/20-mm criterion. The goals-per-episode metric counts targets completed under the strict criterion before reset.

As shown in Table II, directly training on the full goal range rarely produces successful 6D reaching. The curriculum raises strict success from 6.2% to 59.5% on full-range targets and from 3.4% to 55.3% on maximum-change targets. Near success and the number of completed goals improve consistently as well. Thus, progressively expanding the goal range is important for learning large coupled translations and rotations.

TABLE III

GRASP-MAINTENANCE REWARD ABLATION.

Metric	Without grasp reward	**With grasp reward**
Grasp-wrench quality Q	0.032 ± 0.002	$\mathbf{0.045 \pm 0.005}$
Minimum force margin	0.077 ± 0.003	$\mathbf{0.107 \pm 0.010}$
Minimum torque margin	0.036 ± 0.002	$\mathbf{0.047 \pm 0.004}$
Episode success rate (%)	53.7 ± 2.5	$\mathbf{59.1 \pm 2.5}$

3) Effect of the Grasp-Maintenance Reward: We isolate the grasp-maintenance reward by setting only the weight of r_t^{grasp} to zero. We independently train each configuration three times. For each paired comparison, we average the metrics over equal-length training windows after both policies have reached the full 30-mm/180° goal range, and report the mean and sample standard deviation across the three runs.

We focus on four complementary metrics. Grasp-wrench quality Q is the smooth minimum over the six bidirectional force and torque margins defined in Eq. (8). The minimum force and torque margins are the weakest bidirectional coverage values over the three force axes and three torque axes, respectively. Episode success rate is the fraction of completed training episodes in which at least one target is reached. Higher values are better for all four metrics.

As summarized in Table III, the grasp-maintenance reward increases quality by approximately 41%, the minimum force margin by 39%, and the minimum torque margin by 31% on average. All three grasp metrics improve in every run. Episode success rate also rises from 53.7% to 59.1%, a mean gain of 5.4 percentage points. Thus, encouraging the policy to preserve balanced wrench resistance improves the resulting grasp without sacrificing pose-reaching performance.

C. Real-World Capability Evaluation

We evaluate the complete visual-feedback system on hardware across object geometries and wrist orientations, under external disturbances, and during continued reaching without resets.

1) Reaching Across Object Geometries: We evaluate the real-world system on four objects with distinct manipulation geometries: Cube, Hexagonal Prism, Square Bifrustum, and Hammer. The first three are drawn from the three training shape families and sizes and share one geometry-conditioned policy jointly trained on all nine shape–size combinations. They differ in symmetry and surface structure while having comparable scales. The Hammer, measuring $215 \times 50 \times 33.6$ mm, further tests the framework on an object with a substantially larger aspect ratio, moment arm, and translational workspace. Owing to its elongated geometry, the Hammer uses an extended translation curriculum of up to 100 mm, while the maximum rotation remains 180°.

Continuous random goal reaching. For each object, we conduct uninterrupted rollouts in which 6D targets are sampled randomly from its feasible workspace. Once the position and orientation errors remain within their tolerances for the required dwell time, a new target is issued without resetting the hand–object system. In each rollout, POISE completed at least five successive goals without resetting the hand–object system. The experiment therefore evaluates not only whether the policy can reach an individual pose, but also whether it retains sufficient control of the object for subsequent reaching. Figure 6 shows one representative reached pose for each object geometry.

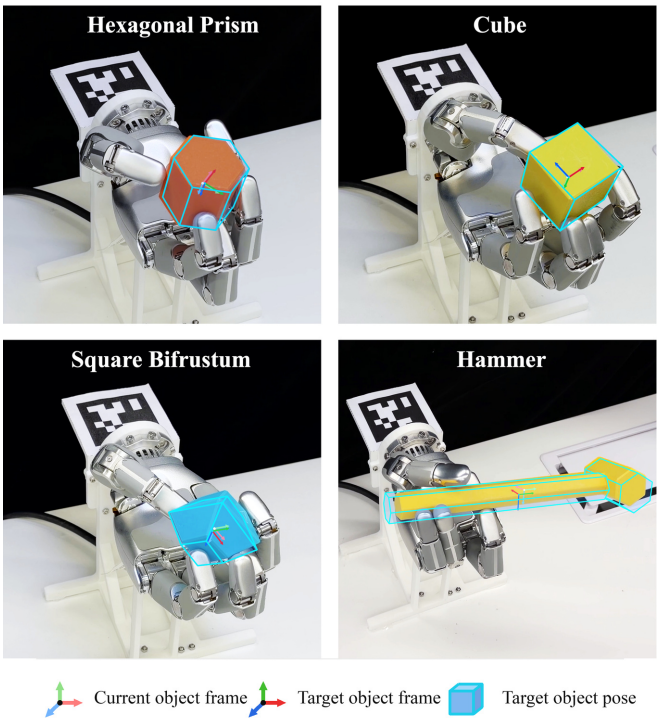

Fig. 6. Representative snapshots after successful 6D pose reaching for the Hexagonal Prism, Cube, Square Bifrustum, and Hammer. The cyan outline denotes the commanded object pose, and the coordinate frames indicate the current and target poses.

Reaching user-specified targets. We further test whether the system can follow explicitly specified goals rather than only targets drawn from the random sampler. For the Hexagonal Prism, the commanded sequence moves the object to different in-hand positions while successively orienting different prism faces upward. For the Hammer, we prescribe full 6D targets involving substantial translation and changes in its principal-axis direction. In the representative sequences, POISE reaches four Hexagonal Prism targets with changes of up to 35 mm and 100°, and seven Hammer targets with changes of up to 93 mm and 180°. The corresponding mean times to reach a target are 3.0 s and 2.7 s. In both cases, each new target is issued from the attained state without resetting the hand or object. Representative reaching sequences are shown in Fig. 1.

2) Reaching Under Different Wrist Orientations: We deploy the same policy with the wrist fixed in three orientations, thereby changing gravity in the palm frame and reducing the passive support available from the palm. Nevertheless, the policy sustains continued reaching in every configuration. In the representative rollouts shown from left to right in Fig. 8, it completes 7, 5, and 6 successive goals without reset, with mean reaching times of 5.1, 5.5, and 7.1 s. Across the three conditions, the per-target minimum error averages 5.3–7.4 mm in position and 3.2°–4.4° in orientation, based on the

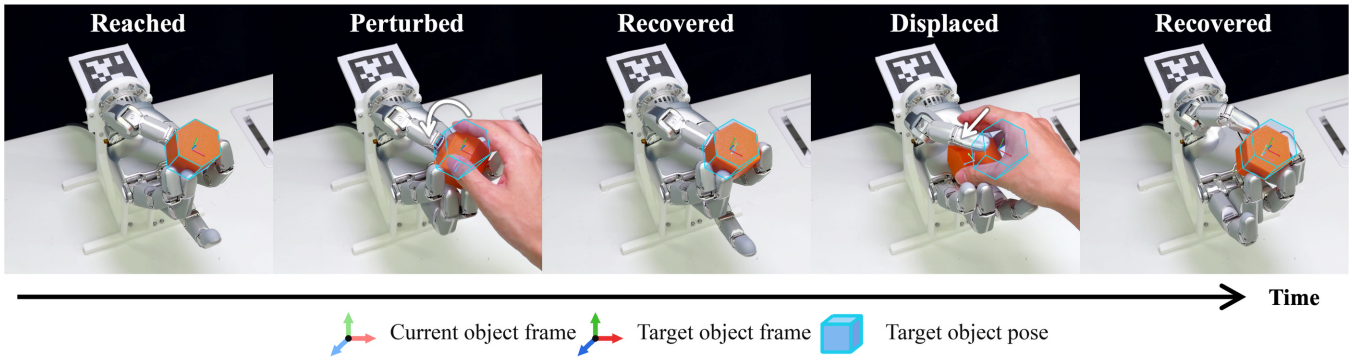

Fig. 7. Closed-loop recovery from two consecutive disturbances on hardware. After reaching the target, the object is first perturbed in orientation and then displaced more substantially, changing its contacts. The policy returns it to the unchanged target after each disturbance.

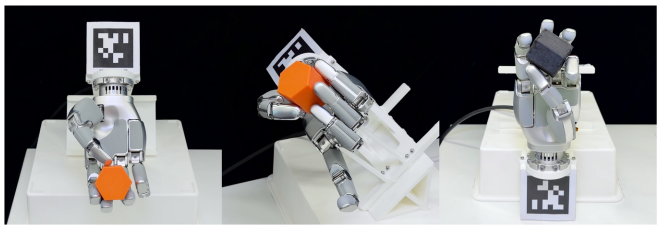

Fig. 8. Representative real-world reaches under three wrist orientations with different gravity directions in the palm frame.

TABLE IV

REAL-WORLD REACHING SUCCESS WITH AND WITHOUT THE GRASP REWARD.

Metric	Without grasp reward	**With grasp reward**
Complete sequences	2/10	**8/10**
Planned targets completed	8/30	**27/30**
Mean targets reached	0.8	**2.7**
Reach time (s)	6.8	**4.9**
Position error (mm)	4.8	**6.4**
Orientation error (°)	4.7	**5.9**

online pose estimates. These results demonstrate continued reaching under different gravity directions using one policy.

3) Closed-Loop Recovery from Disturbances: We further evaluate whether the policy can recover from unexpected disturbances during real-world operation. After the object reaches the commanded pose, an experimenter perturbs it twice while keeping the target unchanged. As shown in Fig. 7, the first perturbation primarily changes the object orientation, whereas the second produces a larger displacement and a change in the contact configuration. In both cases, the policy responds to the updated visual pose estimate, reorganizes the contacts, and returns the object to the commanded pose. This sequence demonstrates closed-loop recovery from externally induced pose and contact deviations.

4) Hardware Ablation of the Grasp-Maintenance Reward: We deploy the two policies from Sec. V-B.3 in 10 trials each. Starting from the same grasp, each policy attempts the same sequence of three targets without reset; completing all three defines sequence success. Reach time and steady-state errors are reported as medians over reached targets, with errors averaged over the final 20 policy steps.

As shown in Table IV, the reward raises sequence success from $2/10$ to $8/10$ and target success from 26.7% to 90.0%, while maintaining comparable pose accuracy. Without the reward, progressive contact loss often led to object drops. The full policy instead tended to retain more distributed multi-finger contacts, improving the reliability of continued reaching.

VI. CONCLUSION AND FUTURE WORK

We presented POISE, a sim-to-real RL framework for reaching palm-relative 6D object poses through finger motions alone. Our results show that diverse stable-grasp initialization improves generalization to unseen grasp configurations and recovery after contact loss, while the adaptive goal curriculum and grasp-preserving reward enable stable, large-range pose reaching. Together, these findings establish palm-relative $\mathrm{SE}(3)$ reaching as a useful primitive for functional in-hand reconfiguration, with the potential to support tool use and other downstream tasks as a step toward general-purpose dexterous manipulation.

Limitations and future work. Two limitations remain. First, POISE specifies the object target pose but not a desired hand or contact configuration. The policy may therefore use effective but unnatural hand postures. Conditioning on both object and grasp targets could produce more natural, task-appropriate motions. Second, real-world performance still lags that in simulation because of imperfect contact modeling and errors in 6D pose tracking. Future work will reduce this gap by incorporating online point-cloud observations through observation distillation, adding tactile feedback, and adapting from real interactions, for example by learning residual dynamics.

REFERENCES

[1] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[2] J. Yin, H. Qi, J. Malik, J. Pikul, M. Yim, and T. Hellebrekers, "Learning in-hand translation using tactile skin with shear and normal force sensing," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 5850–5856.

[3] A. S. Morgan, K. Hang, B. Wen, K. Bekris, and A. M. Dollar, "Complex in-hand manipulation via compliance-enabled finger gaiting and multi-modal planning," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4821–4828, 2022.

[4] H. Qi, A. Kumar, R. Calandra, Y. Ma, and J. Malik, "In-hand object rotation via rapid motor adaptation," in *Proceedings of the 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205. PMLR, 2023, pp. 1722–1732.

[5] T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, and P. Agrawal, "Visual dexterity: In-hand reorientation of novel and complex object shapes," *Science Robotics*, vol. 8, no. 84, p. eadc9244, 2023.

[6] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, V. Kumar, and W. Zaremba, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," *arXiv preprint arXiv:1802.09464*, 2018.

[7] H. J. Charlesworth and G. Montana, "Solving challenging dexterous manipulation tasks with trajectory optimisation and reinforcement learning," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 1496–1506.

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[9] L. Han and J. C. Trinkle, "Dextrous manipulation by rolling and finger gaiting," in *Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, vol. 1, 1998, pp. 730–735.

[10] N. C. Dafle, R. Holladay, and A. Rodriguez, "In-hand manipulation via motion cones," in *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[11] B. Sundaralingam and T. Hermans, "Relaxed-rigidity constraints: Kinematic trajectory optimization and collision avoidance for in-grasp manipulation," *Autonomous Robots*, vol. 43, no. 2, pp. 469–483, 2019.

[12] P. Chanrungmaneekul, K. Ren, J. T. Grace, A. M. Dollar, and K. Hang, "Non-parametric self-identification and model predictive control of dexterous in-hand manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 8743–8750.

[13] Y. Jiang, M. Yu, X. Zhu, M. Tomizuka, and X. Li, "Contact-implicit model predictive control for dexterous in-hand manipulation: A long-horizon and robust approach," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 5260–5266.

[14] H. J. T. Suh, T. Pang, T. Zhao, and R. Tedrake, "Dexterous contact-rich manipulation via the contact trust region," *The International Journal of Robotics Research*, vol. 45, no. 9, pp. 1418–1454, 2026.

[15] G. Khandate, S. Shang, E. T. Chang, T. L. Saidi, J. Adams, and M. Ciocarlie, "Sampling-based exploration for reinforcement learning of dexterous manipulation," in *Proceedings of Robotics: Science and Systems*, 2023.

[16] Z.-H. Yin, C. Wang, L. Pineda, F. Hogan, K. Bodduluri, A. Sharma, P. Lancaster, I. Prasad, M. Kalakrishnan, J. Malik, M. Lambeta, T. Wu, P. Abbeel, and M. Mukadam, "DexterityGen: Foundation controller for unprecedented dexterity," 2025. [Online]. Available: https://arxiv.org/abs/2502.04307

[17] H. Zhang, J. Ferchow, J. Song, and M. Meboldt, "UniCross: Unified cross-skill dexterous manipulation synthesis," *arXiv preprint arXiv:2607.28198*, 2026.

[18] L. Pei, T. Wu, H. Zhang, P. Luo, and J. Song, "Assembling two parts in one hand," in *Conference on Robot Learning*, 2026, to appear.

[19] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, *et al.*, "Solving Rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[20] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 164. PMLR, 2022, pp. 297–307.

[21] A. Handa, A. Allshire, V. Makoviychuk, A. Petrenko, R. Singh, J. Liu, D. Makoviichuk, K. Van Wyk, A. Zhurkevich, B. Sundaralingam, *et al.*, "DeXtreme: Transfer of agile in-hand manipulation from simulation to reality," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5977–5984.

[22] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, "General in-hand object rotation with vision and touch," in *Proceedings of the 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 229. PMLR, 2023, pp. 2549–2564.

[23] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, "Rotating without seeing: Towards in-hand dexterity through touch," in *Proceedings of Robotics: Science and Systems*, 2023.

[24] M. Yang, C. Lu, A. Church, Y. Lin, C. Ford, H. Li, E. Psomopoulou, D. A. W. Barton, and N. F. Lepora, "AnyRotate: Gravity-invariant in-hand object rotation with sim-to-real touch," 2024. [Online]. Available: https://arxiv.org/abs/2405.07391

[25] X. Liu, H. Wang, and L. Yi, "DexNDM: Closing the reality gap for dexterous in-hand rotation via joint-wise neural dynamics model," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=80vjyj5o7l

[26] A. Bhardwaj, M. Wilder-Smith, M. Mittal, V. Patil, and M. Hutter, "ViserDex: Visual sim-to-real for robust dexterous in-hand reorientation," in *Proceedings of Robotics: Science and Systems*, 2026.

[27] J. Yin, Z. Zhao, X. Tan, Y. Liu, C. Wang, and X. Gu, "WM-Craftnet: World synesthesia model for generalizable and robust dexterous in-hand manipulation," in *Conference on Robot Learning*, 2026.

[28] W. Huang, I. Mordatch, P. Abbeel, and D. Pathak, "Generalization in dexterous manipulation via geometry-aware multi-task learning," arXiv preprint arXiv:2111.03062, 2021.

[29] J. Pitz, L. Röstel, L. Sievers, D. Burschka, and B. Bäuml, "Learning a shape-conditioned agent for purely tactile in-hand manipulation of various objects," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 13 112–13 119.

[30] K. Kedia, T. G. W. Lum, J. Bohg, and C. K. Liu, "SimToolReal: An object-centric policy for zero-shot dexterous tool manipulation," arXiv preprint arXiv:2602.16863, 2026.

[31] T. G. W. Lum, K. Kedia, C. K. Liu, and J. Bohg, "Play2Perfect: What matters in dexterous play pretraining for precise assembly?" arXiv preprint arXiv:2606.26428, 2026.

[32] Y. Kuang, S. Park, K. Fragkiadaki, and S. Tulsiani, "Dex4D: Task-agnostic point track policy for sim-to-real dexterous manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2602.15828

[33] X. Liu, J. Adalibieke, Q. Han, Y. Qin, and L. Yi, "DexTrack: Towards generalizable neural tracking control for dexterous manipulation from human references," in *International Conference on Learning Representations*, 2025.

[34] K. Li, P. Li, T. Liu, Y. Li, and S. Huang, "ManipTrans: Efficient dexterous bimanual manipulation transfer via residual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 6991–7003.

[35] Y. Wang, R. Yu, H. W. Tsui, X. Lin, H. Zhang, Q. Zhao, K. Fan, M. Li, J. Song, J. Wang, Q. Chen, and P. Tan, "Learning generalizable hand-object tracking from synthetic demonstrations," *arXiv preprint arXiv:2512.19583*, 2025.

[36] P. Li, Z. Chen, Y. Wu, P. Wei, Y. Li, T. Wang, J. Shi, M. Yu, B. Jia, S.-C. Zhu, T. Liu, and S. Huang, "Towards human-level dexterous teleoperation," *arXiv preprint arXiv:2607.11481*, 2026.

[37] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "DexGraspNet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 359–11 366.

[38] Z.-H. Yin and P. Abbeel, "Lightning Grasp: High performance procedural grasp synthesis with contact fields," *arXiv preprint arXiv:2511.07418*, 2025.

[39] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 4331–4340.

[40] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Munoz, X. Yao, R. Zurbrügg, N. Rudin, *et al.*, "Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.

[41] Sharpa, "Sharpa Wave," accessed: Aug. 27, 2026. [Online]. Available: https://www.sharpa.com/pages/wave

[42] RealSense, "RealSense D435," accessed: Aug. 27, 2026. [Online]. Available: https://www.realsenseai.com/cn/products/stereo-depth-camera-d435/

[43] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 868–17 879.