

Regular Paper

Tactile Reinforcement Learning for Multi-Directional Rotation of a Tiny Sphere on Fingertips

Hongjing Zhu¹ , Haoliang Xu¹ , and Qiang Li^{1,*} 

¹ School of Artificial Intelligence, Shenzhen Technology University, 3002 Lantian Road, Pingshan District, Shenzhen, Guangdong 518118, China (e-mail: emiyarei33@gmail.com; haolianghsu@gmail.com; liqiang1@sztu.edu.cn)

* Corresponding author: Qiang Li (e-mail: liqiang1@sztu.edu.cn).

Abstract: Dexterous in-hand manipulation often relies on multiple fingers or palm support to stabilize an object. Manipulating a tiny object only between the index fingertip and thumb is more challenging because small changes at either contact can quickly lead to slip or object loss. We study continuous multi-directional rotation of a tiny sphere held between two tactile fingertips. To address the partial observability of object motion and contact evolution, we use a teacher-student framework. A privileged teacher is trained with reinforcement learning using object and contact information, while a gravity curriculum gradually increases the task difficulty. The learned behavior is transferred through interactive policy distillation to a student that uses a history of proprioceptive and dual-fingertip tactile observations. The high-resolution tactile maps are summarized into a compact representation for history-based control. Simulation results demonstrate sustained multi-directional manipulation by the privileged teacher, with the distilled student retaining this behavior using only proprioceptive and dual-fingertip tactile histories.

Keywords: dexterous manipulation; tactile sensing; reinforcement learning; in-hand manipulation

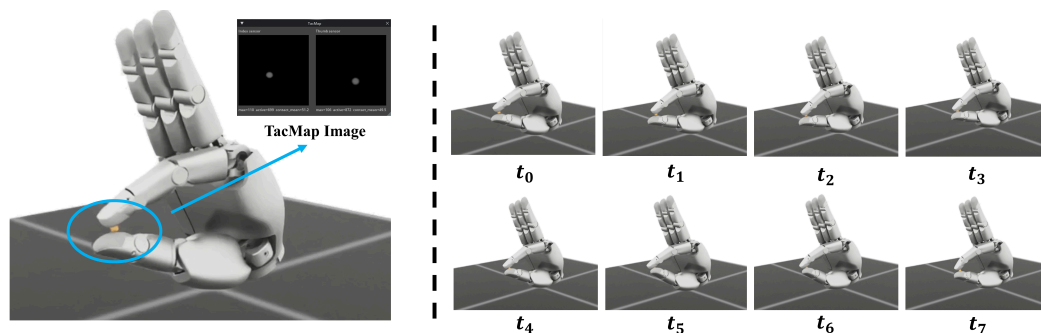


Fig. 1. Overview of the tiny-sphere rotation task. A 5 mm-diameter sphere is held between tactile sensors on the index fingertip and thumb. The left panel shows the dual-fingertip tactile depth maps, and the right panel illustrates a sequence of sustained task motions.

1. INTRODUCTION

Dexterous in-hand manipulation requires a robotic hand to coordinate object motion with changes in contact while maintaining a stable grasp. Friction, contact compliance, and intermittent contact transitions make these dynamics difficult to represent with

fixed analytical assumptions. Early work in this research line first developed a hierarchical feedback controller for unknown objects. The controller used estimated fingertip contact positions and forces to generate local object motion and was evaluated under noisy sensing in physics simulation [1]. This approach was subsequently extended to larger manipulation ranges by coordinating continuous local manipulation with discrete finger-gait transitions through a finite-state machine [2]. A later study introduced online surface exploration during regrasp- ing, enabling a free finger to search for locally favorable contact points while accounting for both manipulability and grasp stability [3]. More recently, rein-

Manuscript received 4 September 2026; revised 29 September 2026; accepted 29 September 2026; and published 30 September 2026. This research was supported by the Natural Science Foundation of TopTalent of SZTU (Grant No. GDRC202411). The material in this paper was not presented at any conference and has not been published previously. This paper was recommended for publication by Associate Editor Hangil Joe.



forcement learning has provided a data-driven route to object reorientation [4], Rubik’s Cube manipulation [5], continuous object rotation [6], and dynamic pen spinning [7]. Related learning-based approaches have also been applied to robotic bagging with compact state representations and action primitives [8], and to flexible-manipulator tracking with explicit friction compensation [9]. Together, these studies trace a progression from feedback-based local control and structured regrasp coordination to learned policies capable of coordinating multiple joints and contacts over extended manipulation sequences.

The remaining challenge is to construct a learnable control problem when the manipulation is governed by a small number of local contacts. Existing systems use proprioceptive histories [6], tactile sensing [10, 11], vision, or multimodal observations [12]. Many of their tasks also benefit from multiple fingers, intermediate links, or palm support. Fingertip-only manipulation requires the object to be stabilized and moved through small, continuously changing contact regions. Similar joint configurations can correspond to different contact locations and object motions, so proprioception alone provides an incomplete description of the interaction. This challenge becomes more pronounced for objects that fit within the opposing fingertip contact surfaces.

The motion of such objects is governed primarily by local fingertip interactions. Small changes in contact location can cause uncontrolled motion, contact loss, or object drop. To study this problem, we consider continuous multi-directional rotation of a small sphere held between two tactile fingertips. We use TacMap [13] to generate tactile images in simulation. It maps contact on each curved fingertip surface to a two-dimensional image, where each pixel encodes local indentation. The resulting dual-fingertip tactile images are shown in the left panel of Fig. 1. The directional command is specified in the local image coordinates of the thumb sensor. The complete task formulation is presented in Sec. 2.

Teacher–student learning has been used to transfer visual representations across heterogeneous network architectures [14] and to adapt robot policies using privileged training information [15]. Following the latter principle, we address partial observability by separating task learning from state inference. A privileged teacher first learns the control behavior from object states, physical parameters, and simulator contact information. A student then learns from the teacher using proprioception, command variables, compact TacMap-derived tactile features, and observation history. Proximal Policy Optimization (PPO) [16] trains the teacher, a gravity curriculum based on curriculum-learning principles [17]

supports early exploration, and DAgger-inspired interactive policy distillation [18] exposes the student to states induced by its evolving policy. The methodological contribution lies in connecting these established learning tools to a task model centered on two local tactile contacts. The interface defines directional commands and contact-center progress in the thumb TacMap coordinates. It supports transfer of the learned behavior to a history-based student.

The main contributions are as follows:

- We formulate a contact-centric, command-conditioned model for continuous multi-directional manipulation between two tactile fingertips.
- We develop a privileged-to-non-privileged learning interface that combines latent teacher supervision, temporal observation encoding and interactive policy distillation.
- Simulation evaluation shows that the student retains most of the teacher’s angular-velocity projection score and bilateral-contact ratio under matched conditions.

2. CONTACT-CENTRIC TASK MODELING

We formulate manipulation between two tactile fingertips as a contact-centric, command-conditioned control problem, as illustrated in Fig. 1. The index fingertip and thumb provide the active contacts, while the remaining fingers and palm provide no support.

2.1. Task Process

We formulate the task as a command-conditioned control problem with partial observability. The underlying physical process is described by a Markov decision process

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma), \quad (1)$$

where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\mathcal{P}$ represents the hand-object transition dynamics, r is the reward function, and γ is the discount factor.

At policy step t , the environment state is written in compact form using its principal physical variables:

$$\mathbf{s}_t = (\mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{p}_t^o, \mathbf{R}_t^o, \mathbf{v}_t^o, \boldsymbol{\omega}_t^o, \mathbf{c}_t), \quad (2)$$

where $\mathbf{q}_t$ and $\dot{\mathbf{q}}_t$ are the positions and velocities of the controlled joints. The object position, orientation, linear velocity, and angular velocity are denoted by $\mathbf{p}_t^o$, $\mathbf{R}_t^o$, $\mathbf{v}_t^o$, and $\boldsymbol{\omega}_t^o$, respectively. The variable $\mathbf{c}_t$ represents the contact states between the object and the

two fingertips. Controller states and task memory are included implicitly in $\mathbf{s}_t$.

Given an action $\mathbf{a}_t \in \mathcal{A}$, the state evolves according to

$$\mathbf{s}_{t+1} \sim \mathcal{P}(\cdot | \mathbf{s}_t, \mathbf{a}_t). \quad (3)$$

The task additionally provides a directional command $\mathbf{g}_t \in \mathcal{G}$ that specifies the desired direction of motion. Because the reward terms measure the outcome of each action, the instantaneous reward is conditioned on the current state, action, next state, and command:

$$r_t = r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, \mathbf{g}_t). \quad (4)$$

The objective is to learn a policy that maximizes the expected discounted return

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]. \quad (5)$$

2.2. Action Space

For the Sharpa Wave right hand used in this work [19], the policy controls four joints of the index finger and five joints of the thumb, while the remaining fingers are maintained at fixed configurations. At each policy step, the policy outputs a nine-dimensional joint-position increment

$$\mathbf{a}_t = [a_t^1 \quad \dots \quad a_t^9]^\top. \quad (6)$$

The action space is defined by the joint-wise lower and upper bounds

$$\mathcal{A} = \{\mathbf{a} \in \mathbb{R}^9 \mid \mathbf{a}_{\min} \leq \mathbf{a} \leq \mathbf{a}_{\max}\}, \quad (7)$$

where the inequalities are applied element-wise. The action updates the joint-position targets according to

$$\mathbf{q}_{t+1}^{\text{tar}} = \text{clip}(\mathbf{q}_t^{\text{tar}} + \mathbf{a}_t, \mathbf{q}_{\min}, \mathbf{q}_{\max}), \quad (8)$$

where $\mathbf{q}_{\min}$ and $\mathbf{q}_{\max}$ denote the joint-position limits. The resulting joint targets are tracked by a position controller.

Object rotation emerges from coordinated changes in the index and thumb configurations, which generate tangential motion while maintaining bilateral contact.

2.3. Multi-Directional Rolling Command

Figure 2 shows how the two-dimensional thumb-TacMap coordinates define the desired tangential motion of the thumb-object contact region. The axes u and v are fixed to the thumb-mounted sensor. At policy step t , the command is represented by a unit vector

$$\mathbf{g}_t = \begin{bmatrix} g_t^u \\ g_t^v \end{bmatrix}, \quad \|\mathbf{g}_t\|_2 = 1. \quad (9)$$

The command specifies the desired tangential motion of the thumb-object contact region in the TacMap image. Under the TacMap parameterization, each image-space direction (g_t^u, g_t^v) maps to a tangential direction on the curved thumb surface at the current contact. We use the following eight canonical commands, separated by 45° :

$$\mathcal{G} = \left\{ \begin{bmatrix} \cos(k\pi/4) \\ \sin(k\pi/4) \end{bmatrix} \mid k \in \{0, \dots, 7\} \right\}. \quad (10)$$

During training, a command is randomly sampled from $\mathcal{G}$ and remains active for 20 policy steps, corresponding to 1 s at the 20 Hz policy frequency.

The reward combines command progress, bilateral contact, slip suppression, action smoothness, and joint-limit regularization.

2.4. Task Objective

The task aims to move the thumb contact center along the commanded TacMap direction while maintaining bilateral contact and limiting relative tangential slip. The reward is defined as

$$r_t = \lambda_d r_t^{\text{cmd}} + \lambda_c r_t^{\text{con}} - \lambda_s r_t^{\text{slip}} - \lambda_a r_t^{\text{rate}} - \lambda_l r_t^{\text{lim}} + r_t^{\text{fail}}, \quad (11)$$

where the coefficients set the relative contributions of directional progress, bilateral contact, slip suppression, action smoothness, and joint-limit avoidance. Table A1 lists their values.

Let $\mathbf{p}_t^T$ denote the thumb contact-center position expressed in meters along the local u and v directions for reward calculation. The tactile-center features supplied to the policy use normalized image coordinates. We define $c_t^I, c_t^T \in \{0, 1\}$ as the contact indicators of the index fingertip and thumb, respectively, and

$$b_t = c_t^I c_t^T. \quad (12)$$

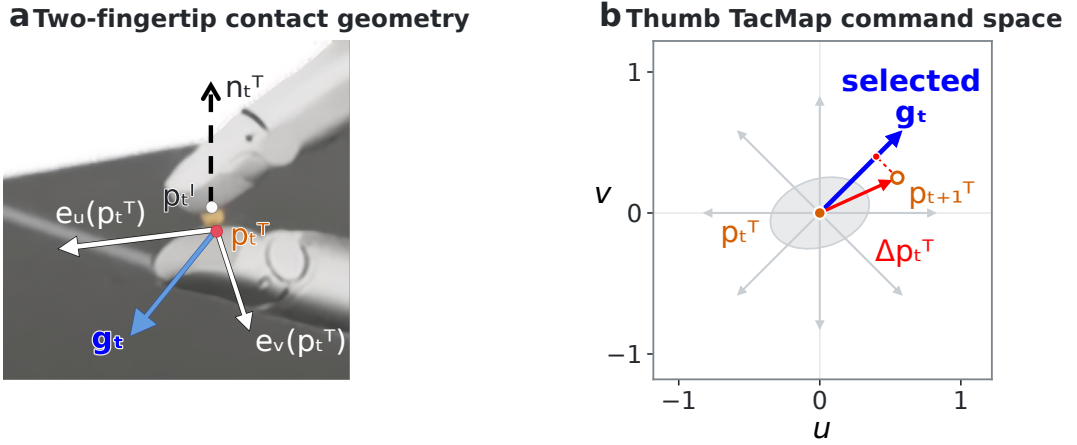


Fig. 2. Local contact geometry and multi-directional rolling commands. (a) The index and thumb contact centers are denoted by $\mathbf{p}_t^I$ and $\mathbf{p}_t^T$. At the thumb contact, $\mathbf{e}_u(\mathbf{p}_t^T)$ and $\mathbf{e}_v(\mathbf{p}_t^T)$ span the local tangent plane, $\mathbf{n}_t^T$ is the surface normal, and $\mathbf{g}_t$ is the selected tangential command. (b) The corresponding command in the thumb TacMap plane. The shaded ellipse represents the contact patch around the current contact center, placed at the origin for illustration. The eight directions in $\mathcal{G}$ are shown with the selected $\mathbf{g}_t$ highlighted in blue and the remaining directions in gray. The red arrow denotes the contact-center displacement $\Delta \mathbf{p}_t^T = \mathbf{p}_{t+1}^T - \mathbf{p}_t^T$, and the red dotted line indicates its projection onto $\mathbf{g}_t$.

Thus, $b_t = 1$ only when both fingertips are simultaneously in contact with the object, and $b_t = 0$ when either contact is absent. The command-progress and bilateral-contact terms use b_{t+1} ; each fingertip's slip contribution uses its own contact indicator c_{t+1}^f . The command-following reward is defined as

$$r_t^{\text{cmd}} = b_{t+1} \text{clip} \left(\frac{\mathbf{g}_t^\top (\mathbf{p}_{t+1}^T - \mathbf{p}_t^T)}{d_{\text{ref}}}, -1, 1 \right), \quad (13)$$

where $d_{\text{ref}} = 0.0025 \text{ m}$ (2.5 mm) is the reference contact displacement. The inner product measures the thumb contact-center displacement along the commanded TacMap direction. A positive value indicates motion along the command, whereas a negative value indicates motion in the opposite direction. Multiplication by b_{t+1} ensures that directional progress is rewarded only when both contacts remain valid after the action. This term measures the task-level progress requested in the local tactile frame.

To encourage bilateral contact, we define

$$r_t^{\text{con}} = b_{t+1}, \quad (14)$$

which provides a positive reward when the binary contact indicators for the index fingertip and thumb are both active.

An auxiliary contact-motion penalty discourages command progress achieved through excessive relative tangential motion. Let $\mathbf{x}_t^f$ be the object-fingertip

contact point for fingertip $f \in \{I, T\}$, and let $\mathbf{n}_t^f$ be the corresponding unit contact normal. The velocities of the sphere and fingertip surface at this point are

$$\mathbf{v}_{o,t}^f = \mathbf{v}_t^o + \boldsymbol{\omega}_t^o \times (\mathbf{x}_t^f - \mathbf{p}_t^o), \quad (15)$$

$$\mathbf{v}_{h,t}^f = \mathbf{v}_{\text{tip},t}^f + \boldsymbol{\omega}_{\text{tip},t}^f \times (\mathbf{x}_t^f - \mathbf{p}_{\text{tip},t}^f). \quad (16)$$

The relative tangential slip velocity is obtained by removing the normal component of their relative velocity:

$$\mathbf{v}_{\text{slip},t}^f = (\mathbf{I} - \mathbf{n}_t^f (\mathbf{n}_t^f)^\top) (\mathbf{v}_{o,t}^f - \mathbf{v}_{h,t}^f). \quad (17)$$

We then define the normalized bilateral slip penalty as

$$r_t^{\text{slip}} = \frac{1}{2} \sum_{f \in \{I, T\}} c_{t+1}^f \text{clip} \left(\frac{\|\mathbf{v}_{\text{slip},t+1}^f\|^2}{v_{\text{ref}}}, 0, 1 \right), \quad (18)$$

where $v_{\text{ref}} = d_{\text{ref}}/\Delta t = 0.05 \text{ m/s}$ for the policy interval $\Delta t = 0.05 \text{ s}$. The penalty is small when the sphere and fingertip surface move with similar tangential velocities, and larger when substantial relative sliding occurs. Simulator contact kinematics provide the relative velocities used during teacher training and evaluation.

The action-rate penalty regularizes changes between consecutive joint commands:

$$r_t^{\text{rate}} = \|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2^2. \quad (19)$$

The joint-limit penalty is defined as

$$r_t^{\text{lim}} = \frac{1}{9} \sum_{j=1}^9 \mathbb{I} \left[q_t^j \notin \mathcal{Q}_j^{\text{safe}} \right], \quad (20)$$

where $\mathcal{Q}_j^{\text{safe}}$ excludes the outer 5% of the valid range of joint j .

Finally, manipulation failures receive a terminal penalty:

$$r_t^{\text{fail}} = \begin{cases} -100, & \text{object drop or invalid state,} \\ -20, & \text{contact loss for three steps,} \\ 0, & \text{otherwise.} \end{cases} \quad (21)$$

Here, contact loss refers to either fingertip losing contact with the sphere for three consecutive policy steps. When termination conditions overlap, invalid numerical or physical states are handled first, object drop second, and three-step contact loss third. The first applicable condition determines the terminal penalty and termination label, preventing a single transition from receiving multiple failure penalties.

Together, these terms reward command progress while bilateral contact is preserved, penalize relative sliding at both contacts, and regularize the joint motion. Each episode begins from a validated two-fingertip grasp and continues across successive commands until manipulation failure or the maximum episode horizon is reached.

3. LEARNING METHOD

We use a teacher–student framework, as illustrated in Fig. 3. A teacher is first trained in simulation using privileged information, including the object state, physical parameters, and contact measurements. The trained teacher then guides a student that receives proprioceptive and tactile observations. The student uses a history encoder to estimate the teacher’s privileged latent representation from a short observation history. The final student operates from the non-privileged observation interface.

3.1. Tactile Representation and Observation Design

Each fingertip produces a single-channel TacMap image encoding local indentation. TacMap has a native resolution of 240×240 ; the configured sampling stride of eight produces a 30×30 image for feature extraction. Let $\mathbf{M}_t^f \in [0, 1]^{H \times W}$ denote this image after

Table 1. Shared policy observation.

Observations	Dimension
Joint position	9
Joint velocity	9
Joint target	9
Previous action	9
Direction and phase	4
Dual-fingertip tactile representation	12
Total	52

dividing its 8-bit output by 255, with $H = W = 30$. We summarize each image with a compact tactile feature

$$\mathbf{t}_t^f = [c_t^{f,\text{map}}, u_t^f, v_t^f, A_t^f, \bar{d}_t^f, d_{t,\text{max}}^f]^\top \in \mathbb{R}^6, \quad (22)$$

for each fingertip $f \in \{I, T\}$. The active pixel set is $\Omega_t^f = \{(u, v) \mid M_t^f(u, v) > d_{\text{th}}\}$, with $d_{\text{th}} = 1/255$. The tactile contact indicator is $c_t^{f,\text{map}} = \mathbb{I}[|\Omega_t^f| > 0]$. For an active map, (u_t^f, v_t^f) is the intensity-weighted centroid in normalized image coordinates, and $A_t^f = |\Omega_t^f|/(HW)$ is pixel coverage. The quantities $\bar{d}_t^f$ and $d_{t,\text{max}}^f$ are the mean and maximum normalized intensity over Ω_t^f . Image coordinates lie in $[-1, 1]$, while coverage and intensity statistics lie in $[0, 1]$. All six features are zero when the active pixel set is empty. Concatenating the index-fingertip and thumb features yields the 12-dimensional tactile observation.

The shared policy observation groups the sensing, command, and control variables provided to both teacher and student:

$$\mathbf{o}_t = [\mathbf{q}_t; \dot{\mathbf{q}}_t; \mathbf{q}_t^{\text{tar}}; \mathbf{a}_{t-1}; \mathbf{g}_t; \sin \phi_t; \cos \phi_t; \mathbf{t}_t^I; \mathbf{t}_t^T], \quad \mathbf{o}_t \in \mathbb{R}^{52}. \quad (23)$$

Here, $\mathbf{q}_t$, $\dot{\mathbf{q}}_t$, and $\mathbf{q}_t^{\text{tar}}$ describe the controlled joints; $\mathbf{a}_{t-1}$ is the previous action; $\mathbf{g}_t$ and ϕ_t encode the current command; and $\mathbf{t}_t^I$ and $\mathbf{t}_t^T$ are the tactile features of the index fingertip and thumb. The observation is summarized in Table 1. Joint positions, velocities, and targets are normalized by their physical ranges.

The privileged observation $\mathbf{p}_t \in \mathbb{R}^{34}$ contains the additional object, physical, contact, and episode variables used by the teacher, as summarized in Table 2.

The 13-dimensional object state contains the relative position, quaternion orientation, linear velocity, and angular velocity, with dimensions 3, 4, 3, and 3,

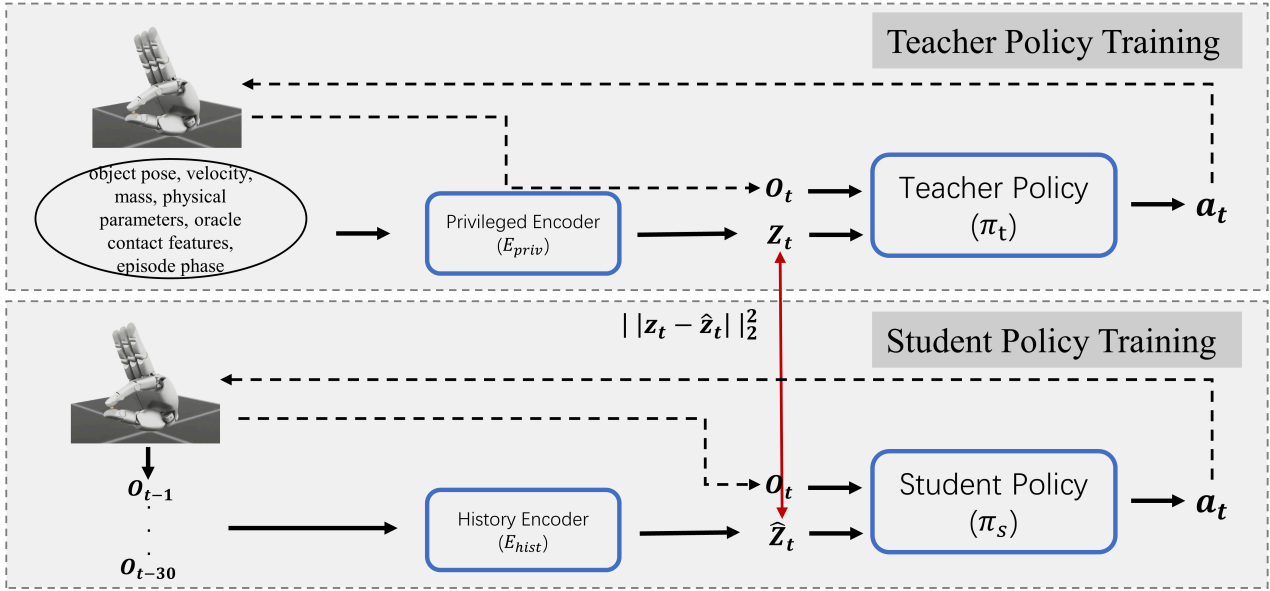


Fig. 3. Teacher-student learning framework. The teacher encodes privileged simulation information into a latent state for policy learning. During distillation, the history encoder predicts this latent state from a history of shared policy observations and is trained with teacher supervision. In the diagram, z_t corresponds to $\mathbf{z}_t^T$ and $\hat{z}_t$ to $\mathbf{z}_t^S$. Solid arrows show policy data flow, dashed arrows show environment-interaction paths, and the red arrow shows latent matching.

Table 2. Privileged observation used by the teacher.

Observations	Dimension
Object pose and velocity	13
Object physical parameters	12
Dual-fingertip contact features	6
Contact stiffness and damping	2
Episode phase	1
Total	34

respectively. The 12-dimensional physical-parameter vector describes the randomized object geometry, inertial properties, friction, and contact parameters. Each fingertip contributes three contact features: contact validity, normal force, and tangential slip speed from the simulator. The remaining entries contain contact stiffness, contact damping, and the current episode phase.

3.2. Privileged Teacher Policy Learning

The teacher provides a stable reference policy for the subsequent student-learning stage. It is optimized with PPO [16], an on-policy reinforcement-

learning algorithm that repeatedly collects trajectories, estimates which actions produced better-than-expected outcomes, and updates the policy while limiting the size of each update. This update constraint helps stabilize exploration in the contact-sensitive task.

The teacher policy receives the shared policy observation together with a latent representation of the privileged information. We use an MLP encoder E_{priv} to encode the privileged observation:

$$\mathbf{z}_t^T = E_{\text{priv}}(\mathbf{p}_t), \quad \mathbf{z}_t^T \in \mathbb{R}^{32}. \quad (24)$$

The resulting latent summarizes the object state, physical properties, and contact information used during teacher training. The latent is concatenated with the shared policy observation and passed to the teacher actor:

$$\mathbf{a}_t \sim \pi_T(\cdot \mid [\mathbf{o}_t; \mathbf{z}_t^T]). \quad (25)$$

The latent representation provides a compact interface between privileged information and the control policy.

3.3. Student Policy Learning

The student learns to reproduce the teacher’s behavior from the shared policy observations defined

in Sec. 3.1. To capture temporal changes in object motion and fingertip contact, the student forms a history from the previous $K = 30$ observations:

$$\mathbf{H}_t = [\mathbf{o}_{t-K}, \dots, \mathbf{o}_{t-1}], \quad K = 30. \quad (26)$$

At the policy frequency of 20 Hz, this history covers the previous 1.5 s of proprioceptive and tactile observations. The current observation $\mathbf{o}_t$ remains a separate actor input. A history encoder E_{hist} maps $\mathbf{H}_t$ to a 32-dimensional estimate of the teacher’s privileged latent representation:

$$\mathbf{z}_t^S = E_{\text{hist}}(\mathbf{H}_t), \quad \mathbf{z}_t^S \in \mathbb{R}^{32}. \quad (27)$$

The history encoder uses three causal one-dimensional convolutional layers with 64, 64, and 128 output channels and kernel sizes of 5, 3, and 3, respectively. Each layer applies left padding to preserve the sequence length and is followed by an ELU activation. Temporal average pooling and a linear projection produce the 32-dimensional latent estimate from the 30×52 observation history.

The student uses the same action interface as the teacher and supplies the history-based latent estimate:

$$\mathbf{a}_t \sim \pi_S(\cdot \mid [\mathbf{o}_t; \mathbf{z}_t^S]). \quad (28)$$

The student actor is initialized from the trained teacher actor. During distillation, the teacher actor and privileged encoder remain frozen and provide fixed action and latent targets. The student therefore learns to infer from observation history the task-relevant information represented by the teacher latent.

We train the student through DAGger-inspired interactive policy distillation [18]. Ordinary behavior cloning trains a student on states visited by the teacher, but small student errors can move the system to states that are underrepresented in the teacher data. DAGger addresses this distribution shift by collecting supervision on states visited by the evolving student. In our procedure, the teacher and student receive observations from the same environment state and independently predict actions $\mathbf{a}_t^T$ and $\mathbf{a}_t^S$. For each parallel environment, the action applied to the simulated hand is selected as

$$\mathbf{a}_t^{\text{exec}} = \begin{cases} \mathbf{a}_t^T, & \text{with probability } \beta_k, \\ \mathbf{a}_t^S, & \text{with probability } 1 - \beta_k. \end{cases} \quad (29)$$

The selected action is executed through the same control interface used during teacher training, advancing the simulator to the next state. Regardless of which action is executed, the teacher latent and action distribution are retained as supervision for the visited state. The mixing probability β_k decreases linearly from one to zero over the first 100 epochs, gradually shifting data collection from teacher-controlled trajectories to states induced entirely by the student.

For each sampled observation history, the student is optimized using latent and action-matching objectives:

$$\mathcal{L}_{\text{latent}} = \frac{1}{32} \|\mathbf{z}_t^S - \mathbf{z}_t^T\|_2^2, \quad (30)$$

$$\mathcal{L}_{\text{action}} = D_{\text{KL}}(\pi_T \parallel \pi_S), \quad (31)$$

which are combined as

$$\mathcal{L}_S = \lambda_z \mathcal{L}_{\text{latent}} + \lambda_\pi \mathcal{L}_{\text{action}}. \quad (32)$$

Both losses are averaged over the minibatch. Their weights are $\lambda_z = 1$ throughout distillation and $\lambda_\pi = 0$ for the first 30 epochs, followed by $\lambda_\pi = 1$. The two losses are jointly applied for the remaining 120 epochs. The interactive procedure therefore supplies teacher supervision on states visited by the evolving student, including states reached after accumulated control errors. The resulting student operates from proprioceptive and TacMap-derived tactile observations.

4. SIMULATION EVALUATION

4.1. Experimental Setup

Simulation and task setup. All experiments are conducted in NVIDIA Isaac Sim [20] using the Sharpa Wave right hand [19]. Teacher training and student distillation are performed on a single NVIDIA GeForce RTX 4090 GPU. The policy controls four joints of the index finger and five joints of the thumb, while the remaining fingers remain fixed. A 5 mm-diameter sphere is held between the tactile index fingertip and thumb.

The physics simulation runs at 240 Hz with a control decimation of 12, yielding a simulated policy frequency of 20 Hz. Each episode lasts for at most 400 policy steps, corresponding to 20 s. A directional command is maintained for 20 policy steps, or 1 s. A complete evaluation episode contains 20 command segments and 19 direction switches. The task requires continuous responses to changing commands while maintaining the grasp throughout the episode.

Each episode starts from a two-fingertip grasp sampled from a precomputed cache. During cache

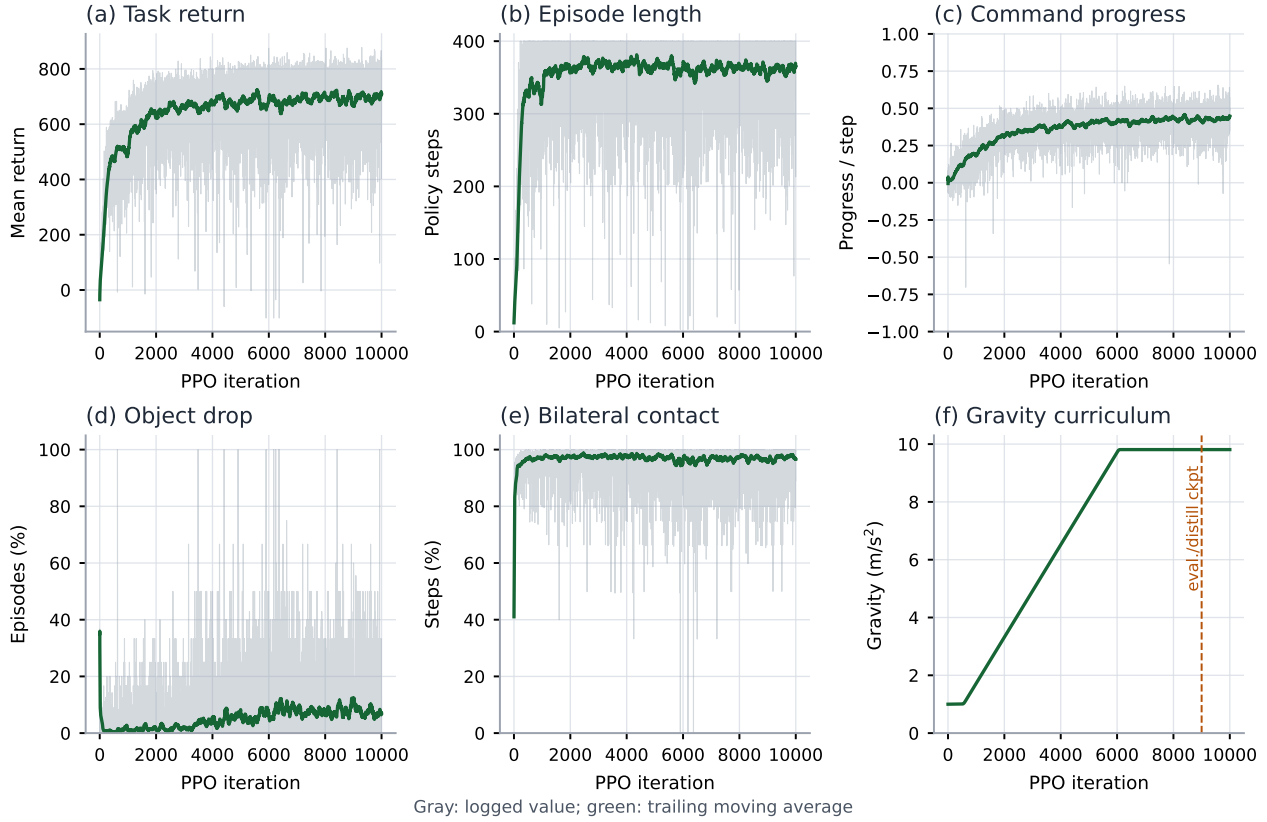


Fig. 4. Teacher training curves over 10,000 PPO iterations. The six panels report task return, episode length, command progress, object-drop rate, bilateral-contact ratio, and the iteration-based gravity curriculum. Gray traces show logged values, and green traces provide a smoothed visual guide to the training trend. The checkpoint at iteration 9,000 is used for student distillation and matched evaluation.

collection at 1.0 m/s^2 , zero actions are applied and states with bilateral contact and no termination are retained. At reset, the cached joint configuration and sphere pose are restored, and joint and object velocities are set to zero. Evaluation proceeds at 9.81 m/s^2 without a separate pre-evaluation settling phase. Episodes follow the termination criteria defined in Sec. 2.4 and the maximum duration specified above.

During training, commands are uniformly sampled from the eight directions defined in the thumb TacMap and are selected independently of the current contact location. During evaluation, the policy follows a fixed sequence that repeatedly covers all eight directions, enabling a controlled comparison of the rotation and contact metrics reported below. The object-drop rate and bilateral-contact ratio quantify object retention. The evaluation covers the same discrete command space used for training; interpolation to unseen directions remains outside the present study.

Teacher training. Curriculum learning organizes training from easier conditions to harder ones so that

a policy can acquire useful behavior before facing the full task difficulty [17]. We instantiate this principle through gravity. Early rollouts use reduced gravity to give exploratory finger motions more time to preserve the object, and the gravity magnitude is then increased according to a fixed iteration schedule. The curriculum changes the training condition only. All student distillation and fixed evaluation use standard gravity.

The privileged teacher is trained with PPO for 10,000 iterations using the task reward defined in Sec. 2. The PPO settings, including the generalized advantage-estimation (GAE) parameter, are listed in Table A1. The gravity magnitude is 1.0 m/s^2 for iterations $i \leq 500$, increases linearly as $g(i) = 1.0 + 8.81(i - 500)/5500$ for $500 < i < 6000$, and is 9.81 m/s^2 for $i \geq 6000$. Gravity is updated before each rollout according to this fixed schedule, with ongoing episodes continuing across updates.

The teacher checkpoint at iteration 9,000 is used to initialize student distillation and for the matched evaluation. Figure 4 shows the full 10,000-iteration

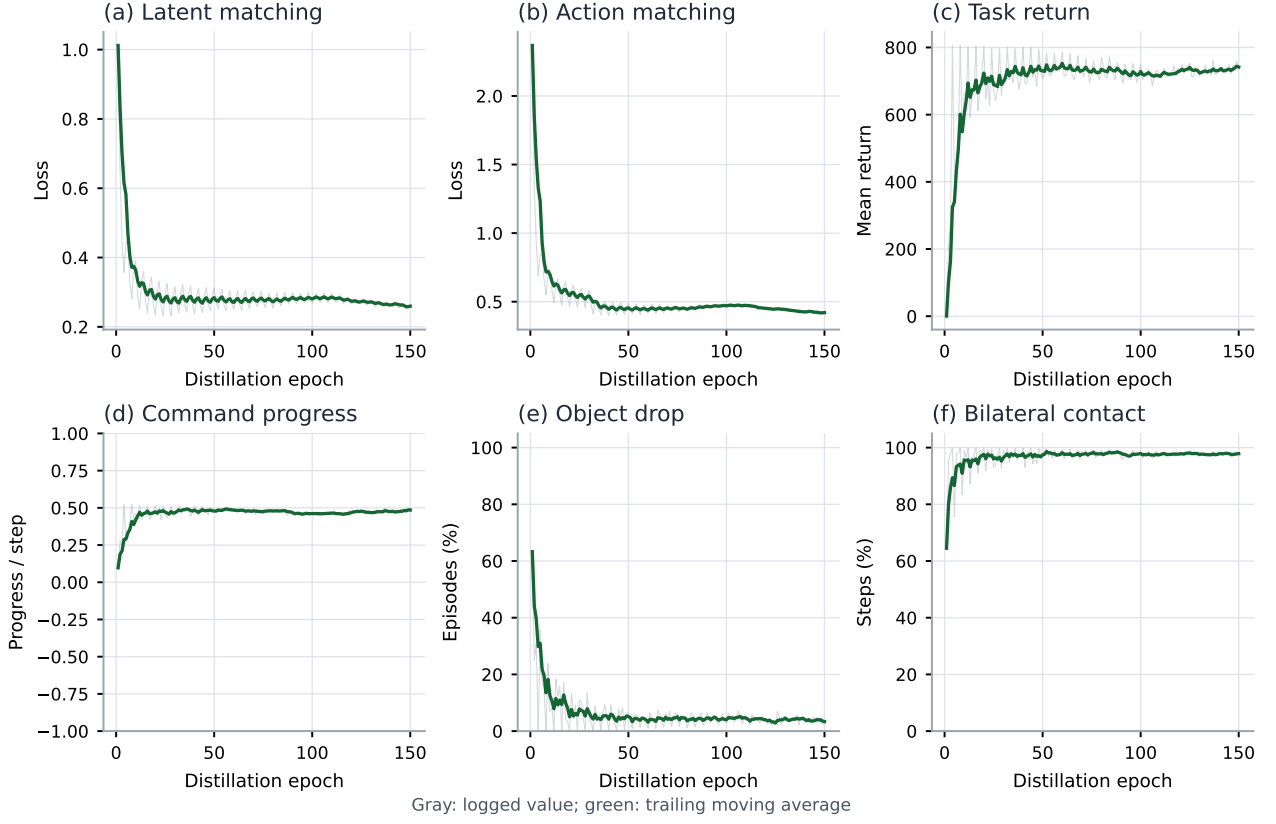


Fig. 5. Student distillation curves over 150 epochs. The six panels report latent-matching loss, action-matching loss, task return, command progress, object-drop rate, and bilateral-contact ratio. Gray traces show logged values, and green traces provide a smoothed visual guide to the training trend. The final student checkpoint is evaluated separately under the matched full-gravity protocol reported in Table 3.

training run.

Student training. The student is trained for 150 epochs following the distillation procedure in Sec. 3.3, with optimization parameters listed in Table A1.

4.2. Framework Performance

The training curves in Fig. 4 and Fig. 5 document optimization of the teacher and student. The rise in the smoothed object-drop rate during approximately iterations 500–6,000 coincides with increasing gravity and is consistent with greater difficulty in retaining the sphere. Variability persists after gravity reaches 9.81 m/s^2 , suggesting that rollout sampling and policy updates may also contribute to the fluctuations.

The teacher checkpoint at iteration 9,000 and the student checkpoint at epoch 150 are evaluated with fixed weights over $N = 256$ paired episodes. Each pair starts from the same cached grasp and follows the same command sequence under deterministic mean actions at standard gravity (9.81 m/s^2). The eight directions cycle through $0^\circ, 90^\circ, 180^\circ, 270^\circ, 45^\circ,$

$135^\circ, 225^\circ,$ and 315° , with the timing specified in Sec. 4.1. Teacher training and student distillation each use one random seed.

Fixed-policy evaluation reports angular-velocity projection scores and direct angular displacements. Command labels follow the thumb TacMap directions defined in Sec. 2. For each label, the evaluator uses the same numerical command components $\tilde{\mathbf{g}}_t = [g_t^u, g_t^v, 0]^\top$ to construct a reference axis in the index-sensor frame. With the contact normal and object angular velocity expressed in this frame, the score is

$$\hat{\mathbf{a}}_{t+1}^{\text{eval}} = \frac{\hat{\mathbf{n}}_{t+1}^I \times \tilde{\mathbf{g}}_t}{\|\hat{\mathbf{n}}_{t+1}^I \times \tilde{\mathbf{g}}_t\|_2}, \quad (33)$$

$$s_t^{\text{ang}} = b_{t+1} \nu_{t+1} \text{clip} \left(\frac{(\hat{\mathbf{a}}_{t+1}^{\text{eval}})^\top \boldsymbol{\omega}_{t+1}^{o,I}}{\omega_{\text{ref}}}, -1, 1 \right), \quad (34)$$

$$\omega_{\text{ref}} = 2 \text{ rad/s}. \quad (35)$$

Here, ν_{t+1} indicates a valid contact normal and projection axis; the score is zero when either is undefined.

Table 3. Performance over 256 paired evaluation episodes per policy. (A) Overall performance and object rotation, reported as mean $\pm$ standard deviation across episodes. (B) Mutually exclusive episode outcomes and the total early-termination fraction. (C) Full-horizon performance in policy-specific, unpaired subsets reaching 400 steps without failure. (D) Student-minus-teacher paired differences with 95% bootstrap confidence intervals.

A. Overall performance and rotation		
Metric	Teacher	Student
Angular-velocity projection score	0.4068 ± 0.1200	0.3834 ± 0.1649
Bilateral-contact ratio (%)	95.88 ± 18.25	88.17 ± 29.39
Cumulative rotation per complete segment (rad)	3.8760 ± 0.5243	3.9274 ± 0.4084
Net rotation per complete segment (rad)	1.7526 ± 0.2519	1.8575 ± 0.2721
Episodes contributing rotation statistics	251	248
Complete command segments	4911	4507
B. Episode outcomes		
Outcome, n/N (%)	Teacher	Student
Object drop	5/256 (1.95)	10/256 (3.91)
Contact loss	9/256 (3.52)	29/256 (11.33)
Invalid state	0/256 (0.00)	0/256 (0.00)
Time limit	242/256 (94.53)	217/256 (84.77)
Total early termination	14/256 (5.47)	39/256 (15.23)
C. Full-horizon survivors		
Metric	Teacher	Student
Episodes	242	217
Angular-velocity projection score	0.4272 ± 0.0845	0.4415 ± 0.0895
Bilateral-contact ratio (%)	99.92 ± 0.19	99.91 ± 0.19
D. Paired differences		
Metric	Student – teacher	95% CI
Angular-velocity projection score	−0.0234	[−0.0409, −0.0062]
Bilateral-contact ratio (percentage points)	−7.71	[−11.10, −4.51]
Object-drop rate (percentage points)	+1.95	[−0.39, 4.30]

For episode e , angular-velocity projection score and bilateral contact are averaged over the scheduled horizon of $T = 400$ steps:

$$P_e = \frac{1}{T} \sum_{t=0}^{T-1} s_{e,t}^{\text{ang}}, \quad (36)$$

$$C_e = \frac{100}{T} \sum_{t=0}^{T-1} b_{e,t+1}. \quad (37)$$

Here, $s_{e,t}^{\text{ang}}$ is the angular-velocity projection score, and $b_{e,t+1}$ is the bilateral-contact indicator recorded

after action t . After early termination, both terms are set to zero for the remainder of the scheduled horizon. These horizon-normalized metrics reflect both performance during execution and the duration of episode survival. We also report each policy’s metrics and sample size for its subset of episodes reaching T without failure.

Episodes are classified by object drop, contact loss, an invalid physical or numerical state, or reaching the time limit. The termination precedence defined in Sec. 2.4 assigns one cause to each episode. The total early-termination fraction includes all episodes

Table 4. Direction-wise angular-velocity projection score under the sequential evaluation protocol. Directions denote the thumb TacMap command labels. Values are mean $\pm$ standard deviation of episode-level direction means. Each direction includes the n episodes that reached it, with the score averaged over its observed steps within each episode.

Direction	Teacher	n	Student	n
0°	0.3383 $\pm$ 0.2126	256	0.2752 $\pm$ 0.2540	256
45°	0.7043 $\pm$ 0.2173	248	0.7283 $\pm$ 0.1765	229
90°	0.2929 $\pm$ 0.1801	251	0.2285 $\pm$ 0.2121	248
135°	0.2822 $\pm$ 0.2101	247	0.2354 $\pm$ 0.2146	228
180°	0.6448 $\pm$ 0.1661	251	0.6621 $\pm$ 0.1742	239
225°	0.6554 $\pm$ 0.1353	246	0.6181 $\pm$ 0.1409	228
270°	0.2259 $\pm$ 0.2023	250	0.3233 $\pm$ 0.2117	234
315°	0.2969 $\pm$ 0.2472	246	0.4236 $\pm$ 0.3157	227

ending before step T . The object-drop rate is $D = 100N_{\text{drop}}/N$.

Direct object rotation is measured from the recorded world-frame orientations $\mathbf{R}_{e,t}^o$. Let

$$d(\mathbf{R}_a, \mathbf{R}_b) = \cos^{-1} \left[\text{clip} \left(\frac{\text{tr}(\mathbf{R}_a^\top \mathbf{R}_b) - 1}{2}, -1, 1 \right) \right]. \quad (38)$$

For a complete command segment starting at step t_s , with $L = 20$ transitions, cumulative and net angular displacement are

$$A_{e,s} = \sum_{j=0}^{L-1} d(\mathbf{R}_{e,t_s+j}^o, \mathbf{R}_{e,t_s+j+1}^o), \quad (39)$$

$$N_{e,s} = d(\mathbf{R}_{e,t_s}^o, \mathbf{R}_{e,t_s+L}^o). \quad (40)$$

Rotation statistics, reported in radians, include episodes with at least one complete segment. Complete-segment values are averaged within each episode before computing the mean and standard deviation across contributing episodes. Cumulative displacement is resolved at the 20 Hz policy sampling rate; net displacement is the shortest start-to-end orientation difference.

For angular-velocity projection score, bilateral-contact ratio and object-drop rate, we report student-minus-teacher paired differences with 95% percentile bootstrap confidence intervals. These intervals are computed from 10,000 resamples of the 256 episode pairs.

Direct orientation measurements showed object rotation during complete command segments for both policies (Table 3A). The student achieved 94.3% of the teacher’s mean angular-velocity projection score and 92.0% of its mean bilateral-contact ratio (paired differences and 95% confidence intervals: Table 3D). Early termination was more frequent for the student, with contact loss accounting for most of its failures. Among full-horizon survivors, mean bilateral-contact ratios exceeded 99.9% for both policies.

Angular-velocity projection score varied across the eight directions under the sequential protocol (Table 4). Both policies had their highest mean score at 45°. The lowest means occurred at 270° for the teacher and 90° for the student. Contact states carry over between successive commands, and early termination reduces exposure to later directions; the direction-wise means therefore describe performance within this sequence.

5. CONCLUSION AND LIMITATIONS

We introduced a tactile fingertip-manipulation task for controlling a small sphere between the index fingertip and thumb under successive local directional commands while maintaining bilateral contact. To address partial observability, we developed a teacher-student framework that combines privileged teacher training with history-based policy distillation. The student estimates the teacher’s privileged latent representation from a history of proprioceptive and tactile observations.

Using proprioceptive and tactile observations, the student retained most of the teacher’s angular-velocity projection score and bilateral-contact ratio under matched simulation evaluation. These results support the feasibility of the proposed task and learning framework within the evaluated simulation setting.

5.1. Limitations

The present study is conducted entirely in simulation. The ray-casting TacMap approximates contact depth over the curved fingertip surface but does not model elastomer deformation, tactile hysteresis, measurement drift, or nonuniform sensor response.

The evaluation focuses on a 5 mm-diameter sphere and a single fingertip geometry. Extending the task to other object scales and shapes will require compatible fingertip geometries and contact configurations.

Evaluation also begins from validated two-fingertip grasps and covers the same eight directions used during training. Initial sphere acquisition, unseen directions, and general $SO(3)$ pose tracking are

not evaluated.

5.2. Future Work

Future work will evaluate the contribution of tactile observations and observation history and extend the task to broader object scales, shapes, and contact configurations. Further evaluation will also consider more realistic tactile sensing and transfer beyond the current simulation setting.

APPENDIX A

A.1. Reward and Optimization Parameters

Table A1. Reward and optimization parameters. Penalty coefficients are positive magnitudes, with minus signs specified in the reward equation. Distillation weights correspond to the loss definitions in Sec. 3.

Parameter	Value
Command-progress weight, λ_d	2.0
Bilateral-contact weight, λ_c	0.5
Slip penalty weight, λ_s	0.2
Action-rate penalty weight, λ_a	0.01
Joint-limit penalty weight, λ_l	0.02
Reference contact displacement, d_{ref}	2.5 mm
Reference slip speed, v_{ref}	0.05 m/s
PPO discount	0.99
GAE parameter	0.95
PPO clip	0.2
PPO learning rate	3×10^{-4}
Distillation learning rate	3×10^{-4}
Latent loss weight, λ_z	1
Action loss weight, λ_π (epochs 1–30)	0
Action loss weight, λ_π (epochs 31–150)	1

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

- [1] Q. Li, R. Haschke, H. Ritter, and B. Bolder, “Simulation results for manipulation of unknown objects in hand,” in *2011 IEEE International Conference on Robotics and Biomimetics*, 2011, pp. 1949–1954.
- [2] Q. Li, M. Meier, R. Haschke, H. Ritter, and B. Bolder, “Object dexterous manipulation in hand based on finite state machine,” in *2012 IEEE International Conference on Mechatronics and Automation*, 2012, pp. 1185–1190.
- [3] Q. Li, R. Haschke, B. Bolder, and H. Ritter, “Grasp point optimization by online exploration of unknown object surface,” in *2012 12th IEEE-RAS International Conference on Humanoid Robots*, 2012, pp. 417–422.
- [4] OpenAI: Marcin Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, 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.
- [5] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang, “Solving Rubik’s Cube with a robot hand,” arXiv preprint, 2019, arXiv:1910.07113.
- [6] 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. [Online]. Available: <https://proceedings.mlr.press/v205/qi23a.html>
- [7] J. Wang, Y. Yuan, H. Che, H. Qi, Y. Ma, J. Malik, and X. Wang, “Lessons from learning to spin ‘Pens’,” in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 3124–3138. [Online]. Available: <https://proceedings.mlr.press/v270/wang25j.html>
- [8] F. Munguia-Galeano, J. Zhu, J. D. Hernández, and Z. Ji, “Learning to bag with a simulation-free reinforcement learning framework for robots,” *IET Cyber-Systems and Robotics*, vol. 6, no. 2, 2024, Art. no. e12113.
- [9] Y. Li, B. Yu, Y. Jian, Y. Qin, and A. An, “A reinforcement learning-based nonlinear trajectory tracking control strategy for space flexible manipulator with LuGre friction compensation,” *IET Cyber-Systems and Robotics*, vol. 8, no. 1, 2026, Art. no. e70040.
- [10] 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*, Daegu, Republic of Korea, Jul. 2023.
- [11] W. Hu, B. Huang, W. W. Lee, S. Yang, Y. Zheng, and Z. Li, “Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing,” *Robotics and Autonomous Systems*, vol. 186, 2025, Art. no. 104904.
- [12] 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. [Online]. Available: <https://proceedings.mlr.press/v229/qi23a.html>
- [13] L. Su, Z. Peng, R. Ren, S. Mao, J. Du, K. Zhang, and X. Zhu, “Tacmap: Bridging the tactile sim-to-real gap via geometry-consistent penetration depth map,” arXiv preprint, 2026, arXiv:2602.21625.

- [14] Z. Wang, S. Han, X. Wang, J. Hao, X. Cao, and B. Zhang, "Big2Small: Learning from masked image modelling with heterogeneous self-supervised knowledge distillation," *IET Cyber-Systems and Robotics*, vol. 6, no. 4, 2024, Art. no. e70002.
- [15] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "RMA: Rapid motor adaptation for legged robots," in *Proceedings of Robotics: Science and Systems*, Virtual, Jul. 2021.
- [16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint, 2017, arXiv:1707.06347.
- [17] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 41–48.
- [18] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. *Proceedings of Machine Learning Research*, vol. 15. PMLR, 2011, pp. 627–635. [Online]. Available: <https://proceedings.mlr.press/v15/ross11a.html>
- [19] Sharpa Robotics, "Wave Dexterous Robotic Hand," Product documentation, 2025, accessed: August 4, 2026. [Online]. Available: <https://www.sharpa.com/pages/wave>
- [20] NVIDIA Corporation, "NVIDIA Isaac Sim," 2025, robotics simulation platform built on NVIDIA Omniverse. [Online]. Available: <https://github.com/isaac-sim/IsaacSim>



Hongjing Zhu is currently pursuing a bachelor's degree in unmanned aerial vehicle systems engineering at Shenzhen Technology University. Her research interests include robot learning and robotic manipulation.



Haoliang Xu received his bachelor's degree in Internet of Things engineering from Shenzhen Technology University in 2025. His research interests include robot learning and robotic manipulation.



Qiang Li received the Ph.D. degree in pattern recognition and intelligent systems from the Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS), Shenyang, China, in 2010. He is currently a Full Professor at Shenzhen Technology University, Shenzhen, China.

From 2009 to 2012, he was a Postdoctoral Researcher with CoR-Lab, Bielefeld University, Bielefeld, Germany. From 2012 to 2018, he was a Senior Scientific Researcher with Bielefeld University. From 2019 to 2020, he was a Principal Scientific Researcher and Project Leader with RoboticsX Lab. He was a project investigator for DEXMAN, funded by the Deutsche Forschungsgemeinschaft (DFG), Bonn, Germany, and worked with the University of Hamburg, Hamburg, Germany. His current research interests include tactile servoing and recognition, sensory-based robotic dexterous manipulation, and robotic calibration and dynamic control. Dr. Li received a stipend from the Honda Research Institute Europe, Offenbach, Germany.