UniBYD: A Unified Framework for Learning Robotic Manipulation Across Embodiments Beyond Imitation of Human Demonstrations

Tingyu Yuan[1,2], Biaoliang Guan[3], Wen Ye[2,4], Ziyan Tian[1,2], Yi Yang[5], Weijie Zhou[1,6], Zhaowen Li[8], Yan Huang[2,4], Peng Wang[1,9], Chaoyang Zhao[1,7*], and Jinqiao Wang[1,2,7*]

[1] Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[4] New Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China
[5] Central South University, Changsha, China
[6] Beijing Jiaotong University, Beijing, China [7] Objecteye inc., Beijing, China
[8] Yinwang Intelligent Technology Co. Ltd., Beijing, China
[9] CasiaHand Robotics Co., Ltd., Beijing, China
yuantingyu2024@ia.ac.cn, chaoyang.zhao@nlpr.ia.ac.cn

Abstract. In embodied intelligence, the embodiment gap between robotic and human hands brings significant challenges for learning from human demonstrations. Although some studies have attempted to bridge this gap using reinforcement learning, they remain confined to merely reproducing human manipulation, resulting in limited task performance. Moreover, current methods struggle to support diverse robotic hand configurations. In this paper, we propose UniBYD, a unified framework that uses a dynamic reinforcement learning algorithm to discover manipulation policies aligned with the robot's physical characteristics. To enable consistent modeling across diverse robotic hand morphologies, UniBYD incorporates a unified morphological representation (UMR). Building on UMR, we design a dynamic PPO with an annealed reward schedule, enabling reinforcement learning to transition from offline-informed imitation of human demonstrations to online-adaptive exploration of policies better adapted to diverse robotic morphologies, thereby going beyond mere imitation of human hands. To address the severe state drift caused by the incapacity of early-stage policies, we design a hybrid Markov-based *shadow engine* that provides fine-grained guidance to anchor the imitation within the expert's manifold. To evaluate UniBYD, we propose UniManip, the first benchmark for cross-embodiment manipulation spanning diverse robotic morphologies. Experiments demonstrate a 44.08% average improvement in success rate over the current state-of-the-art. Our project page is https://zhanheng-creator.github.io/UniBYD/.

* Corresponding authors.

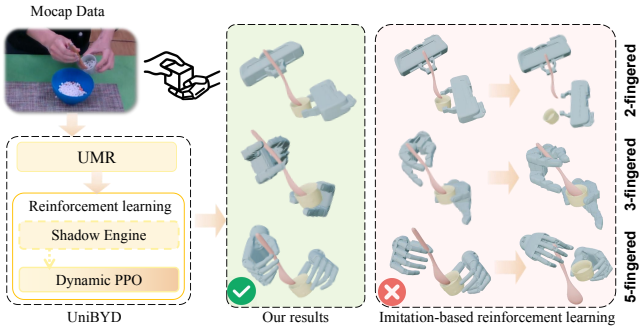

Fig. 1: Leveraging human demonstrations, UniBYD learns manipulation strategies that transcend mere imitation and are tailored to a broad spectrum of robotic hand morphologies.

Keywords: Learning from human demonstrations · Dexterous manipulation · Reinforcement learning · Embodied AI

1 Introduction

Artificial intelligence has advanced rapidly in recent years [37, 38, 42], with remarkable progress in robotics. Within embodied intelligence, learning from human demonstrations [16, 18, 25, 40, 48] has emerged as a dominant paradigm. However, the embodiment gap between human hands and robotic hands of varying morphologies [17] poses significant challenges for this area, which existing studies have yet to effectively address. Notably, most research focuses on anthropomorphic hands, while cross-embodiment generalization to 2/3-fingered grippers remains largely unexplored. For example, retargeting-based methods typically map only the kinematic poses while ignoring critical dynamic information. Existing imitation learning methods remain at merely reproducing human operations [25, 33]. Given the morphological and dynamic discrepancies between human hands and robot hands of varying structures, such as differences in finger count and degrees of freedom, such direct reproduction limits them far below the level demonstrated by humans.

To overcome the limitations of the above approaches, researchers have begun exploring reinforcement learning from human data. However, such methods often struggle to discover strategies that are truly aligned with the robot's own morphology, resulting in suboptimal task performance, especially for non-anthropomorphic hands with significant topological deviations. Some studies, such as ManipTrans [13], employ Proximal Policy Optimization (PPO) [27] in simulated environments, yet their reward functions merely enforce strict time alignment of the robot hand's joint angles with the expert trajectory at every step, which only partially addresses the inherent limitations of imitation learning. As shown on the right of Fig. 1, such methods merely map human-hand

motions onto robotic hands, resulting in low success rates for task execution. Secondly, another line of research attempts to completely eliminate dependence on human demonstrations by defining reward functions centered on object pose errors [3]. This comes at the cost of losing the crucial guidance that human prior provides, making it difficult to approach the high-performance policy regions. These methods tend to fall into local optima. Beyond these limitations, a fundamental challenge for reinforcement learning in this domain is the severe state drift in early training. Due to the lack of effective guidance, even minor deviations from human priors can cause the robot to stray from correct trajectories and terminate episodes prematurely, resulting in extremely limited information gain and failed bootstrapping. Moreover, existing methods exhibit limited generalization, as most are tailored to specific robotic hands and lack a unified framework for adapting human policies to diverse robots [43].

To address these challenges, we introduce a novel paradigm that shifts from rigid motion imitation to morphology-adaptive policy discovery, ensuring manipulation strategies are inherently aligned with the mechanical characteristics of diverse robotic embodiments. According to this paradigm, we introduce UniBYD, a unified reinforcement learning framework designed to acquire morphology-adaptive manipulation policies that transcend mere imitation and generalize across diverse embodiments. Firstly, to realize cross-embodiment generalization, we introduce a unified morphological representation (UMR), which standardizes the state-action space and employs morphology embeddings to enable configuration-aware modeling. Building on UMR, we propose a dynamic PPO mechanism with reward annealing, which enables a smooth transition from offline-informed imitation to online-adaptive exploration. This process guides the model to discover policies better suited to the physical potential of diverse robots. To mitigate the severe state drift in early training, we design a hybrid Markov-based *shadow engine* that provides fine-grained guidance to anchor the robot's execution near expert trajectories, ensuring consistent information gain.

To our knowledge, UniBYD is the first to learn manipulation policies for diverse embodiments from human demonstrations with reinforcement learning. To rigorously evaluate UniBYD, we build UniManip, the first benchmark that spans diverse hand configurations and a wide variety of tasks, providing a standardized measure for manipulation competence. Experimental results show that UniBYD achieves a substantial improvement of up to 44.08% in overall success rate over the current state-of-the-art (SOTA). Our contributions are as follows:

- We propose UniBYD, a unified reinforcement learning framework compatible with various robotic hand types. The framework learns control strategies aligned with the diverse embodiments better.
- We design a dynamic PPO, which facilitates a systematic transition from offline-informed imitation to online-adaptive exploration, enabled by a hybrid Markov-based *shadow engine* that anchors early-stage training to prevent trajectory drift.
- We construct UniManip, the first unified benchmark based on human demonstrations for evaluating robotic manipulation capability across morphologies.

Experiments on UniManip demonstrate that UniBYD significantly outperforms existing SOTA methods, achieving a 44.08% gain in success rate.

2 Related Work

Robotic manipulation learning methods. Achieving human-level dexterity remains a central challenge in robotics [5,6,8,10,11,19,21,29,36,46,47]. Classical approaches such as trajectory optimization [24,30,36] and STOCS [45] compute joint trajectories and contact forces, yet are predominantly offline and thus limited in real-time applicability. Model Predictive Control [1, 9, 10] enables online planning via finite-horizon optimization, but its computational burden hampers real-time deployment. Reinforcement learning optimizes policies through environment interaction [4, 41], supporting the training of complex tasks. However, the exploration space is vast, a challenge UniBYD addresses by using human demonstrations to provide crucial guidance and rapidly focus the policy search.
Learning from Human Demonstrations. Given the high cost of collecting robotic manipulation data, numerous methods have sought to learn from human demonstrations [26,49]. Conventional inverse-kinematics retargeting [14,34] and learning-based retargeting [14,23,28] offer limited performance. Mainstream practice leverages reinforcement learning to acquire robotic manipulation skills from human hand data. Imitation-centric approaches map demonstrations directly onto robots [25]. For example, ManipTrans [13] reproduces human actions on a 5-fingered platform via a universal trajectory imitator and residual modules. However, it fails to surpass the imitation of demonstrations and ignores the embodiment gap. In contrast, goal-centric methods [17], such as DexMachina [22], prioritize task objectives while largely ignoring imitation rewards to encourage exploration, but training is lengthy and convergence is often precarious. Moreover, developing a unified framework that generalizes across diverse robotic hand morphologies remains a critical challenge.
Benchmarks. Existing benchmarks for robotic manipulation have been proposed [12], including Bi-DexHands [2], which assembles dozens of bimanual tasks in simulation, and VTDexManip [15], which evaluates single- and bimanual manipulation with integrated vision and touch. Benchmarks tailored to LfHD have likewise emerged [39]; for instance, EgoDex [7] compiles 829 hours of egocentric manipulation videos and hand-motion data, and DexMachina [22] introduces a benchmark centered on articulated objects. Nevertheless, a comprehensive, unified benchmark spanning 2-, 3-, and 5-fingered hand morphologies across diverse single-hand and bimanual tasks is still lacking. Moreover, existing evaluation protocols are largely one-dimensional, failing to assess manipulation strategies and their alignment with diverse robotic hand embodiments.

3 Method

As shown in Fig. 2, we propose UniBYD, a unified and progressive reinforcement learning framework that learns from humans across various robotic hands.

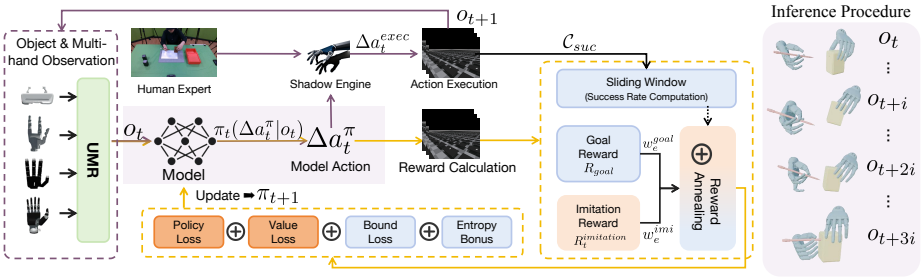

Fig. 2: The framework of UniBYD. UniBYD first encodes diverse hands via UMR. It then employs a dynamic PPO with an annealed reward mechanism, which initially leverages the *shadow engine* for high-fidelity imitation before transitioning to autonomous exploration to discover morphology-aligned policies.

UniBYD aims to discover morphology-aligned policies that transcend mere imitation. During the early training phase, UniBYD blends the action Δa_t^π predicted by the policy network with the expert action Δa_t^E from the demonstration to generate the final executed action used to advance the environment to the next state. In contrast, during the inference phase, UniBYD relies solely on the policy network, directly executing its predicted action Δa_t^π to complete the task.

3.1 Unified Morphological Representation

The challenge of cross-morphology generalization lies in the fundamental differences among robotic hand embodiments. Accordingly, we propose an efficient unified morphological representation.

For robotic hand h, the proprioceptive state s_t^h at step t includes a fixed-dimensional wrist state $s_{base} \in \mathbb{R}^{13}$ and a variable-dimensional joint state s_{joint}^h. The wrist state encodes position, orientation, and velocities, while the joint state contains joint angles and velocities, $q^h, \dot{q}^h \in \mathbb{R}^{D_h}$, where D_h is the hand's degrees of freedom. To avoid the 2π wrap-around issue, joint angles are trigonometrically encoded as $\cos(q^h)$ and $\sin(q^h)$.

Let D_{max} denote the maximum number of joint degrees of freedom. For hands with $D_h < D_{max}$, we apply zero-padding to the variable component s_{joint}^h, thereby elevating its dimensionality to D_{max} and obtaining s_{joint}^{pad}:

$$\begin{aligned} s_{joint}^{pad} = [q^h \oplus \mathbf{0}_{D_{max}-D_h} \oplus \cos(q^h) \oplus \mathbf{0}_{D_{max}-D_h} \oplus \\ \sin(q^h) \oplus \mathbf{0}_{D_{max}-D_h} \oplus \dot{q}^h \oplus \mathbf{0}_{D_{max}-D_h}]. \end{aligned} \tag{1}$$

Moreover, the policy π_θ must be informed of the hand's specific physical attributes. To this end, we extract key static morphological properties from the hand's URDF model, namely D_h, the number of fingers N_{finger}^h, and the number of rigid bodies N_{body}^h. These quantities constitute a static descriptor $v_{morph}^h = [N_{finger}^h, D_h, N_{body}^h]$.

Finally, we concatenate s_{base}, s_{joint}^{pad}, and v_{morph}^{h} to form the observation $o_t^{finger} = s_{base} \oplus s_{joint}^{pad} \oplus v_{morph}^{h}$. By unifying dynamic states and static attributes into a fixed-dimensional representation, UMR enables the policy to adapt to diverse hand morphologies and learn morphology-specific manipulation policies.

3.2 Dynamic Proximal Policy Optimization

With UMR providing a consistent observation space, UniBYD employs a reinforcement learning algorithm integrating a reward annealing mechanism. This mechanism facilitates a systematic transition from offline-informed imitation to online-adaptive exploration, allowing the model to inherit expert knowledge while transcending its limitations through active environmental interaction.

From Offline-Informed Imitation to Online-Adaptive Exploration (1) Imitation Reward ($R^{imitation}$). To achieve precise imitation of offline expert demonstrations, we design a dense, multi-component imitation reward $R^{imitation}$ that, at step t of episode i, quantifies the similarity between the current state s_t and the expert state s_t^E. It is defined as

$$R_t^{imitation} = \sum_{k=1}^{n} w_k \cdot r_k(s_t, s_t^E) - p(\Delta a_t), \tag{2}$$

where r_k denotes the k-th reward component and w_k its corresponding weight. The components primarily encompass discrepancies in wrist pose, linear and angular velocities; fingertip positions and contact forces; joint positions and velocities; and object pose, linear and angular velocities. The term $p(\Delta a_t)$ is a power penalty on actuation. For detailed computations, see Appendix Section 7.

(2) Goal Reward (R^{goal}). To relax the reliance on offline expert demonstrations in dexterous manipulation, we define a sparse goal reward R^{goal} that depends solely on task success. In contrast to the dense $R_t^{imitation}$, this signal is step-agnostic and granted only when the entire episode is successfully completed, thereby conferring a substantial bonus that marks the attempt as productive exploration. R^{goal} is defined as:

$$R^{goal} = \begin{cases} S_{bonus} & \text{if } \mathcal{C}_{suc} = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where S_{bonus} is a fixed-magnitude bonus, and $\mathcal{C}_{suc} \in \{0, 1\}$ denotes the episodic success indicator. In the first phase, $\mathcal{C}_{suc}$ is determined by the discrepancies between the object and fingertip observations and those of the expert, and later solely by the object. More details are provided in Appendix Section 8.

(3) Dynamic Reward Annealing. To enable a smooth transition from imitation to task-centric exploration, we define the total reward R_t as a dynamically weighted sum of the imitation reward $R_t^{imitation}$ and the goal reward R^{goal}:

$$R_t = w_e^{imi} \cdot R_t^{imitation} + w_e^{goal} \cdot R^{goal}, \tag{4}$$

where w_e^{imi} and w_e^{goal} are weighting coefficients for $R_t^{imitation}$ and R^{goal} at training epoch e. The evolution of these weights constitutes a three-stage curriculum, with phase transitions jointly governed by the epoch threshold T_{decay} marking the end of the *shadow engine*, the earliest trigger threshold T_{early}, the recent success rate $\bar{SR}$, and a scaling factor γ_{scale} controlling the transition speed.

To evaluate the current competence of the model in real time, we introduce the recent success rate $\bar{SR}$ as a performance indicator. Specifically, we design a sliding window of size M to aggregate the outcomes of the most recent M episodes, defined as $\bar{SR} = \frac{1}{M}\sum_{j=i-M+1}^{i}\mathcal{C}_{suc}$.

To quantify the progression of task mastery, we introduce a minimal competence threshold δ_{min} (set to 0.2) and a transition factor f:

$$f = \max\left(0, \min\left(1, (\bar{SR} - \delta_{min}) \times \gamma_{scale}\right)\right). \tag{5}$$

The term $(\bar{SR} - \delta_{min})$ primarily serves to prevent w_e^{imi} from decaying under sub-optimal performance or stochastic noise. Simultaneously, it mitigates abrupt weight drops upon triggering the transition mechanism, thereby maintaining training stability during the switchover phase.

Furthermore, we introduce a success-rate trigger threshold δ_{SR} as a criterion to determine whether the model has attained sufficient proficiency to initiate the transition. Building on this, the weights w_e^{imi} and w_e^{goal} are adaptively adjusted according to the following formulations:

$$w_e^{imi} = \begin{cases} w_{max}^{imi} & \text{if } e \le T_{decay} \wedge (e \le T_{early} \vee \bar{SR} \le \delta_{SR}) \\ max(w_{min}^{imi}, w_{max}^{imi} \times (1.0 - f)) & \text{otherwise} \end{cases}, \tag{6}$$

$$w_e^{goal} = \begin{cases} 0 & \text{if } e \le T_{decay} \wedge (e \le T_{early} \vee \bar{SR} \le \delta_{SR}) \\ max(w_{min}^{goal}, w_{max}^{goal} \times f) & \text{otherwise} \end{cases}. \tag{7}$$

This dynamic reward annealing approach implicitly partitions the training process into three stages. In the early imitation-driven alignment phase, while the initial temporal or performance criteria are not yet met, the weights remain fixed to prioritize learning from expert demonstrations and establishing basic manipulation skills. During the hybrid adaptive transition phase, if the model's competence remains limited ($\bar{SR} - \delta_{min} \le 0$), w_e^{imi} is maintained without decay while w_e^{goal} is set to a small value to introduce R^{goal}. Once the model has mastered the manipulation skills, w_e^{imi} enters an annealing mode, decreasing gradually according to changes in $\bar{SR}$ as w_e^{goal} increases. At this stage, the influence of expert data begins to wane, facilitating a "soft-handover" of the guidance focus toward R^{goal}. Finally, in the autonomous exploration phase, the imitation weight reaches its minimum w_{min}^{imi}, exerting a nearly negligible influence. At this point, the policy is predominantly driven by the sparse goal reward, focusing solely on the overall success of the task. This allows the model to move beyond strictly mirroring human manipulation and discover refined control strategies that are better optimized for the robot hand's unique hardware morphology.

Loss Synergy and Counterbalancing To facilitate effective exploration and prevent premature convergence, we incorporate the entropy regularization and boundary loss into the PPO objective, forming a synergy–counterbalance strategy that ensures policy physical feasibility while pursuing optimal performance.

(1) Entropy Regularization. To prevent the policy, particularly during the imitation-guided reinforcement learning phase, from prematurely converging to a suboptimal deterministic solution, we introduce an entropy regularization term, $\mathcal{H}(\pi_\theta(\cdot \mid o_t))$. This term serves as an entropy bonus that encourages sustained exploration. Its coefficient, c_e^{entropy}, follows a linear decay schedule:

$$c_e^{\text{entropy}} = \max\left(0,\ c_{\text{start}}^{\text{entropy}} \left(1 - \frac{e}{T_{\text{entropy\_decay}}}\right)\right), \tag{8}$$

where $c_{\text{start}}^{\text{entropy}}$ is the initial entropy coefficient, and $T_{\text{entropy\_decay}}$ is the decay horizon. This mechanism ensures ample exploration in the early stages of training and gradually reduces exploration thereafter to facilitate convergence.

(2) Bound Loss. While the entropy bonus fosters exploration, the mean of the policy distribution, $\mu_\theta(o_t)$, can readily drift beyond the physical action space. Conventional hard clipping disrupts gradient flow and severely impairs training. To remedy this, we introduce a differentiable soft boundary loss, L^{bound}, which penalizes only clearly out-of-bounds means μ_t:

$$L_t^{bound}(\mu_t) = [\sum_{j=1}^{D_a}(\max(0, \mu_{t,j} - \mu_{bound})^2 + \max(0, -\mu_{t,j} - \mu_{bound})^2)], \tag{9}$$

where D_a denotes the action dimensionality, $\mu_{t,j}$ is the j-th component of the action-mean vector μ_t , and μ_{bound} is a soft boundary threshold.

(3) Dynamic PPO Objective Function. We integrate the foregoing components into the PPO objective, which we minimize:

$$L_t(\theta) = [-L_t^{\text{CLIP}}(\theta) + c_{vf} L_t^{\text{VF}}(\theta) - c_e^{\text{entropy}} \mathcal{H}(\pi_\theta(\cdot \mid o_t)) + c_{\text{bound}} L_t^{\text{bound}}(\mu_t)], \tag{10}$$

where L_t^{CLIP} denotes PPO's clipped surrogate objective, L_t^{VF} is the mean-squared error loss of the value function, and c_{vf} and c_{bound} are their respective weighting coefficients. The entropy bonus term ($-c_e^{\text{entropy}}\mathcal{H}$) and the boundary loss term ($+c_{\text{bound}} L_t^{\text{bound}}$) establish an effective synergy-and-counterbalance: the former fosters broad exploration, while the latter ensures that such exploration remains confined to a physically safe and smooth action space.

3.3 Hybrid Markov-based Shadow Engine

At the outset of training, the policy network π_θ is markedly weak. Within a standard Markov Decision Process (MDP), even a slight action deviation Δa_t can shift the subsequent state s_{t+1}, and the ensuing compounding errors rapidly drive the policy away from meaningful expert trajectories. However, it is difficult for the policy to return to the correct path by relying on the penalties provided

by post-hoc metrics $R^{imitation}$. As a result, this leads to frequent premature episode terminations. Ultimately, the model receives only scarce and weak training signals, impairing overall learning efficiency. Therefore, UniBYD introduces the *shadow engine* to guide the dynamic PPO during the crucial early phase as shown in Fig. 3. It enables fine-grained, efficient imitation by applying dynamic expert guidance to both the robotic hands and the object.

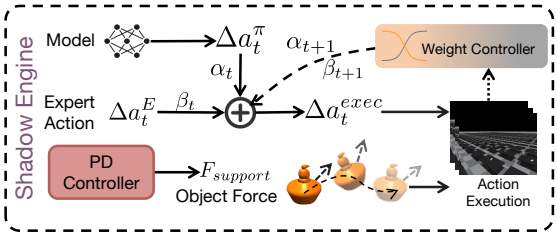

Fig. 3: Overview of action generation and object control in the *shadow engine*. It blends the model-predicted action and the expert-guided action to generate the final executed action Δa_t^{exec}. A PD controller applies an expert object force to guide the object.

Dexterous Hand Control To mitigate compounding errors caused by an initially weak policy, we propose a Hybrid MDP mechanism that unifies discrete pointwise learning with continuous Markov process learning. Specifically, at training step t, the action executed in the simulator, Δa_t^{exec}, is not the raw prediction Δa_t^π from π_θ. It is a dynamically weighted blend of Δa_t^π and the expert demonstration action Δa_t^E.

At step t, conditioned on the current observation o_t, the policy network produces the action Δa_t^π. Concurrently, we retrieve the corresponding expert action Δa_t^E from the demonstration. The executed action Δa_t^{exec} is defined as

$$\Delta a_t^{exec} = \alpha_t \cdot \Delta a_t^\pi + \beta_t \cdot \Delta a_t^E, \tag{11}$$

where α_t is the weight on the policy action Δa_t^π, and β_t is the weight on the expert action Δa_t^E. These weights satisfy $\alpha_t + \beta_t = 1$ throughout training.

We realize dynamic guidance by applying a linear decay curriculum to β_t as a function of the training epoch e:

$$\beta_t = \max\left(0, 1 - \frac{e}{T_{decay}}\right), \tag{12}$$

where T_{decay} is a predefined decay horizon. Correspondingly, α_t increases linearly from 0 to 1. This setup ensures that in the early phase of training $\beta_t \approx 1$ and $\alpha_t \approx 0$, so the model effectively learns each step in isolation without being influenced by the previous step. Once imitation performance is high, $\beta_t = 0$ and $\alpha_t = 1$. At that point $a_t^{exec} = a_t^\pi$, the guidance of the *shadow engine*

disappears, and π_θ must independently handle the full Markov decision process. See Appendix Section 6.1 for details.

Object Control In complex tasks, the object's intrinsic physical properties, such as gravity and inertia, can still precipitate failure. To further ease the initial learning phase, the *shadow engine* also applies a dynamic support force $F_{support}$ directly to the object. We utilize a proportional-derivative (PD) controller to compute $F_{support}$ based on the desired object pose and velocity $(g_t^{obj}, \dot{g}_t^{obj})$ from the expert demonstration. Acting as an invisible hand, $F_{support}$ constrains the object to remain near its target trajectory, preventing drops or catastrophic deviations. The gains of this PD controller (K_p, K_d) are also gradually decayed to zero as training proceeds. For further details, see Appendix Section 6.2.

4 Experiments

4.1 Benchmark

To evaluate generalizable manipulation across diverse hand morphologies, a benchmark must encompass a wide spectrum of tasks involving complex object interactions and diverse functional manipulations. We build UniManip upon OakInk-V2 [44], as it represents one of the largest-scale and most diverse human demonstration datasets, featuring unrivaled synchronization precision and task complexity. Leveraging these high-fidelity motion data, we curate a wide range of unimanual and bimanual tasks and convert them into expert demonstrations tailored to robotic hands with 2, 3, or 5 fingers (see Appendix Section 10 for conversion details and task taxonomy). In total, we assemble 31 task categories to construct UniManip. For 5-finger unimanual and bimanual settings, we select 9 and 6 task categories, respectively. Given that 3-finger and 2-finger hands are ill-suited to bimanual manipulation, we evaluate them on 8 unimanual task categories each. We assess manipulation performance along multiple axes, and specifically define the following metrics:

- **Position Error (PE):** PE is defined as the mean Euclidean distance between the observed and target object positions across time steps in successful episodes. If all test episodes fail, PE is assigned a default value of 3. Lower PE indicates more accurate and stable manipulation.
- **Orientation Error (OE):** OE is defined as the mean minimal rotation angle between the observed and target orientations across time steps in successful episodes. If all test episodes fail, OE is set to a default value of 30. Lower OE reflects more precise orientation control.
- **Success Rate (SR):** The percentage of test episodes classified as successful. Given the process-oriented nature of our tasks, an episode is deemed successful if and only if every time step simultaneously satisfies PE $\leq$ 3 cm and OE $\leq 30°$. Any violation at any time step renders the episode a failure.

Table 1: Comparative results on UniManip. UniBYD consistently outperforms representative baselines across diverse hand morphologies.

Hand Type	Metrics	Retarget	Maniptrans	DexMachina*	UniBYD
2 (1 hand)	SR↑(%)	12.27	✗	✗	**78.13**
	PE↓(cm)	2.81	✗	✗	**0.53**
	OE↓(°)	28.77	✗	✗	**18.74**
	AS↑	4.24	✗	✗	**8.93**
3 (1 hand)	SR↑(%)	4.36	✗	✗	**71.81**
	PE↓(cm)	2.65	✗	✗	**0.89**
	OE↓(°)	29.19	✗	✗	**12.95**
	AS↑	2.48	✗	✗	**9.13**
5 (1 hand)	SR↑(%)	5.68	26.44	✗	**85.67**
	PE↓(cm)	2.89	1.93	✗	**0.77**
	OE↓(°)	29.13	22.28	✗	**10.90**
	AS↑	3.07	5.88	✗	**9.29**
5 (2 hands)	SR↑(%)	2.10	28.75	25.33	**57.67**
	PE↓(cm)	2.96	1.44	1.66	**0.72**
	OE↓(°)	29.56	16.84	14.13	**13.44**
	AS↑	2.58	4.34	4.81	**8.16**

Table 2: Ablation study of UniBYD. The results validate the individual contributions of SE and GR, as well as the incremental gain of LSC, in enhancing performance.

Hand Type	Metrics	base	+SE	+GR	+GR +LSC	UniBYD
2 (1 hand)	SR↑(%)	24.94	67.06	56.19	61.50	**78.13**
	PE↓(cm)	2.66	0.54	1.28	1.32	**0.53**
	OE↓(°)	27.51	19.89	21.31	21.58	**18.74**
	AS↑	5.63	7.56	7.99	8.36	**8.93**
3 (1 hand)	SR↑(%)	19.56	51.06	65.13	67.63	**71.81**
	PE↓(cm)	2.45	0.95	0.93	**0.87**	0.89
	OE↓(°)	26.09	14.67	14.06	**12.85**	12.95
	AS↑	5.42	6.87	8.52	8.78	**9.13**
5 (1 hand)	SR↑(%)	13.56	55.11	63.17	65.39	**85.67**
	PE↓(cm)	2.46	1.24	0.87	0.79	**0.77**
	OE↓(°)	27.12	17.29	13.22	13.23	**10.90**
	AS↑	4.87	6.33	8.29	8.64	**9.29**
5 (2 hands)	SR↑(%)	33.46	43.33	47.01	54.17	**57.67**
	PE↓(cm)	1.46	1.43	1.48	1.36	**0.72**
	OE↓(°)	15.35	17.64	17.17	14.17	**13.44**
	AS↑	3.20	3.55	6.27	6.73	**8.16**

– **Adaptation Score (AS):** This metric aims to quantify how well a manipulation policy aligns with the robot's hardware morphology beyond binary success. We employ a joint evaluation framework, incorporating a suite of heterogeneous large models (including Gemini 2.5 Pro, GPT-5.2, and Qwen2.5-VL-72B) alongside 100 human volunteers as expert evaluators to provide a composite score from 0 to 10. LLMs and humans are blind to success labels to ensure independent assessment of manipulation quality. Please refer to Appendix Section 10.3 for calculation details and prompts. Crucially, AS serves as a qualitative complement to our primary physical metric.

4.2 Experimental Setup

Implementation details. Simulations were conducted in Isaac Gym [20], with 4,096 parallel environments during training. We configured S_{bonus} to 20, T_{decay} to 150, T_{early} to 100, γ_{scale} to 2, δ_{min} to 0.2, and δ_{SR} to 0.8. Additional parameters are detailed in Appendix Section 9.1. For each task in UniManip, we executed 1,000 trials per task in simulation.

Experimental equipment. All experiments were conducted on servers equipped with an NVIDIA GeForce RTX 4090 GPU and an Intel Core i9-14900K CPU. We evaluated our methods on the Franka 2-fingered gripper, the xArm 2-fingered gripper, the CASIA Hand-G 3-fingered dexterous hand [35], the Inspire 5-fingered dexterous hand, and the OHand^{TM} dexterous hand.

4.3 Comparison with SOTA

To comprehensively evaluate UniBYD, we compare it with three representative baselines: a classical retargeting method based on optimization-based inverse kinematics, and two current SOTA methods for dexterous manipulation, ManipTrans [13] and DexMachina* [22]. Since existing SOTA methods do not support 2/3-fingered tasks, and there is a lack of comparative approaches for learning 2/3-fingered dexterous manipulation from human demonstrations, we employ retargeting as the primary baseline for such embodiments. To ensure a fair comparison, we reimplement DexMachina within our simulator, denoted as *DexMachina** throughout this paper. To maintain consistency, all methods share the same experimental configuration, the implementation details of which are provided in Appendix Section 9. The comparative results are reported in Tab. 1, where ✗ denotes that the method does not support that hand type.

The results unequivocally demonstrate that UniBYD surpasses all baselines across every evaluation dimension, achieving a 44.08% improvement in success rate over ManipTrans (SOTA). This underscores UniBYD's superior high-dimensional coordination, whereas imitation-centric ManipTrans is constrained by the embodiment gap and goal-centric DexMachina struggles with complex collaboration without fine-grained guidance. First, UniBYD is the only unified framework that consistently achieves high reliability across all hand morphologies, whereas ManipTrans and DexMachina do not support 2/3-fingered hands. Second, retargeting yields marginal success rates. Even on challenging 5-finger bimanual tasks, UniBYD achieves a 57.67% success rate, outperforming the best baseline by a substantial 28.92% margin. On 5-fingered unimanual tasks, UniBYD achieves an 85.67% SR, outperforming ManipTrans by a 59.23% margin. Simultaneously, UniBYD significantly improves manipulation precision, reducing PE and OE by 60.10% and 51.08% compared to ManipTrans, demonstrating its superior fine-grained control. On 2/3-fingered tasks, UniBYD achieves a high SR of 78.13% and 71.81%, respectively. This underscores its robust cross-embodiment generalization and ability to complete tasks reliably. Finally, UniBYD consistently maintains a superior AS ($\geq$ 8.16), markedly exceeding all baselines (max 5.88). This proves that its learned policies are not only more effective but also more strongly aligned with specific embodiments.

As illustrated in Fig. 4, we visually compare the manipulation strategies learned by UniBYD against those of other methods. ManipTrans seeks to emulate the human tactic of grasping the mug with the three rear fingers. However, these fingers are too wide to pass through the handle, causing the mug to slip and fall. Hampered by sparse reward signals, DexMachina* likewise fails to discover a viable strategy. By contrast, through trial and error, UniBYD identifies that the three robotic fingers are much wider than those of a human. It adapts by using only two fingers (the middle and ring fingers) to grasp the handle, pinching it with the thumb and index finger while bracing the mug with the little finger.

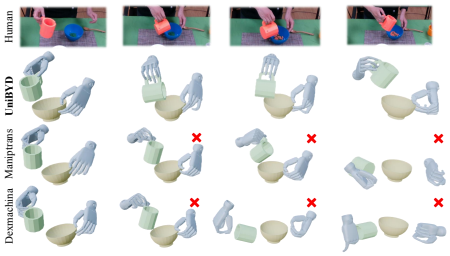

Fig. 4: A visual comparison of the experimental results. UniBYD learns manipulation strategies aligned with the robot's embodiment, thereby successfully completing the task, whereas both ManipTrans and DexMachina* fail.

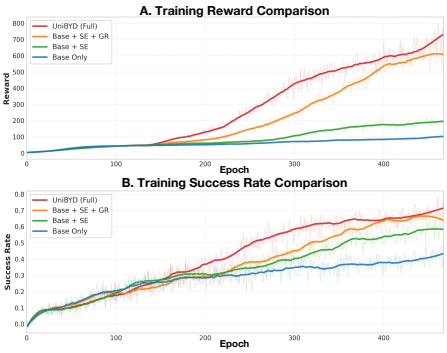

Fig. 5: Training procedure of a representative task.

4.4 Ablation Study

As reported in Tab. 2, we conduct an ablation study built upon the base model. We define the base as an RL approach utilizing only the imitation rewards, functionally equivalent to behavior cloning and serving as the primary baseline for Sec. 4.3. Notably, even base outperforms existing SOTA in bimanual tasks, confirming that our UMR and imitation rewards provide a superior foundation. Compared to the base, +SE (base with *shadow engine*) improves the success rate by 31.26%, validating that the fine-grained expert guidance provided by SE is essential for effective policy initialization. +GR (base with goal reward R^{goal}) achieves a 35.00% SR gain and elevates AS to 7.77. This variant facilitates a seamless transition from offline-data-driven imitation to online goal-driven exploration. This allows the robot to transcend human priors and discover physically optimal postures uniquely tailored to its own morphology. +GR+LSC (base with R^{goal} and loss synergy with counterbalancing) maintains strategic stochasticity, effectively preventing premature convergence to suboptimal imitation. More investigations into other components, including manipulator-specific information(UMR), hyperparameter sensitivity, Boundary Loss, Entropy Loss, and the Reward Annealing schedule are provided in Appendix Section 14.

Fig. 5 illustrates the evolution of reward and success rate in a representative task. Base is unable to explore optimal postures, rapidly settling into a suboptimal policy. In contrast, +SE more accurately reproduces human behavior. Building upon +SE, the introduction of R^{goal} forms +SE+GR. With the fading of expert guidance and under the direction of explicit objectives, this strategy explores postures that lead to higher success rates. Furthermore, combined with LSC (UniBYD), it encourages exploration by maintaining stochasticity in the early phase, eventually converging to a superior manipulation strategy.

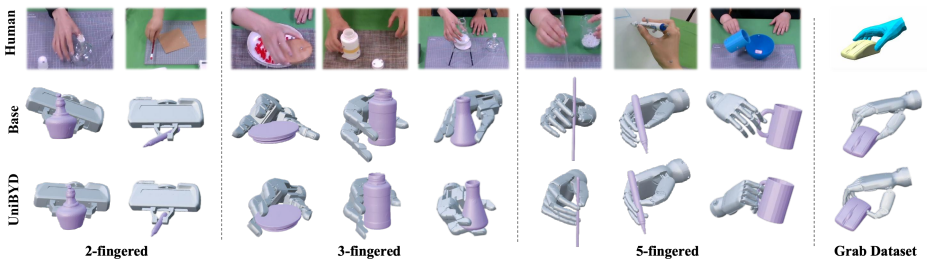

Fig. 6: Comparative experimental results for base and UniBYD.

4.5 Qualitative Analysis: Beyond Mere Imitation

As shown in Fig. 6, UniBYD can discover manipulation strategies aligned with each robot's embodiment. For the 2-fingered gripper, the imitation-only base rigidly reproduces the human's oblique grasp, while UniBYD adopts a more stable grasp oriented perpendicular to the body of the alcohol burner. For the 3-fingered hand, unlike the base that relies solely on two fingertips to pinch the Erlenmeyer flask, UniBYD secures the flask with two fingers while employing the third finger to support the bottom for a more stable manipulation. For the 5-fingered hand, in tasks such as stirring and handwriting, the base imitates a two-fingertip pinch prone to slippage, whereas UniBYD employs the remaining three digits to provide supportive contact forces, enhancing success. To demonstrate the generalizability of UniBYD, we further present results on the GRAB dataset [31] in Fig. 6, where UniBYD achieves a remarkable 98.00% SR and 9.43 AS.

To investigate the progression from imitation to exploration, we compare checkpoints across training epochs, as shown in Fig. 7. By epoch 100, training is dominated by imitation. The policy attempts to manipulate with two fingers but frequently drops it due to limited force control. By epoch 200, the policy begins to relax its reliance on expert data, attempting to grasp the doughnut with the first and third digits while probing the role of the second digit. By epoch 400, it discovers a strategy that pinches the doughnut with two fingers and braces it with the remaining finger. Exploiting the mechanical characteristics, UniBYD observes that operating with the wrist canted sideways enhances success.

4.6 Real-World Experiments

By mapping the dexterous hand's wrist to the robotic arm flange and aligning its degrees of freedom, we conduct experiments on three real-world platforms: the X-Arm 2-fingered hand, the Casia Hand-G 3-fingered dexterous hand, and the OHand^{TM} 5-fingered dexterous hand. To facilitate zero-shot transfer, we employ FoundationPose [32] for hand-object alignment and MoveIt! for trajectory planning. More implementation details are provided in Appendix Section 9.5.

The system achieves success in 26 of 50 trials, 32 of 50 trials, and 35 of 50 trials, respectively. As shown in Fig. 8, for the same task, UniBYD tailors its

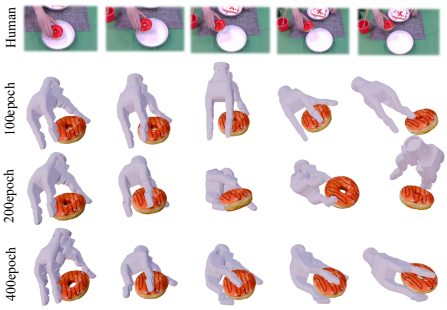

Fig. 7: Results illustrating the progressive evolution of manipulation strategies over the course of training.

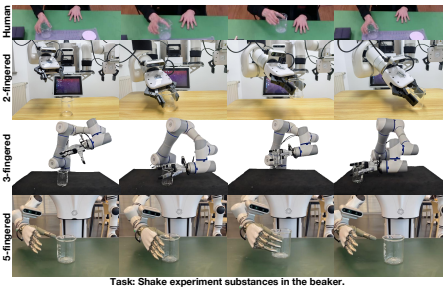

Fig. 8: Experimental results on real-world robotic platforms.

manipulation strategy to the end-effector morphology. With a 2-fingered gripper, it clamps the beaker diagonally, whereas with a 3-fingered hand, it wraps the beaker with all fingers. These results prove that UniBYD transfers effectively to the real world, despite a slight drop in success rate. Failure analysis indicates that failure modes include hand-object collisions (19%), object drops (9%), joint limit triggers (7%), and hand-table collisions (3%).

5 Conclusion

We introduce UniBYD, a unified framework that learns robotic manipulation beyond imitation from human demonstrations. UniBYD achieves a systematic evolution from fine-grained imitation to morphology-adaptive exploration, discovering manipulation strategies that align with each robot's embodiment. Real-world results confirm its superior embodiment alignment and transferability.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62276260 and the National Key Research and Development Program of China (No.2024YFE0210600).

References

1. Ai, B., Tian, S., Shi, H., Wang, Y., Pfaff, T., Tan, C., Christensen, H.I., Su, H., Wu, J., Li, Y.: A review of learning-based dynamics models for robotic manipulation. Science Robotics **10**(106), eadt1497 (2025)
2. Chen, Y., Geng, Y., Zhong, F., Ji, J., Jiang, J., Lu, Z., Dong, H., Yang, Y.: Bidexhands: Towards human-level bimanual dexterous manipulation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(5), 2804–2818 (2023)

3. Chen, Y., Wang, C., Yang, Y., Liu, C.K.: Object-centric dexterous manipulation from human motion data. arXiv preprint arXiv:2411.04005 (2024)
4. Cutler, E., Xing, Y., Cui, T., Zhou, B., van Rijnsoever, K., Hart, B., Valencia, D., Ong, L.V.C., Gee, T., Liarokapis, M., et al.: Benchmarking reinforcement learning methods for dexterous robotic manipulation with a three-fingered gripper. arXiv preprint arXiv:2408.14747 (2024)
5. Gao, X., Yao, K., Khadivar, F., Billard, A.: Real-time motion planning for in-hand manipulation with a multi-fingered hand. CoRR (2023)
6. Gao, X., Yao, K., Khadivar, F., Billard, A.: Enhancing dexterity in confined spaces: real-time motion planning for multifingered in-hand manipulation. IEEE Robotics & Automation Magazine (2024)
7. Hoque, R., Huang, P., Yoon, D.J., Sivapurapu, M., Zhang, J.: Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709 (2025)
8. Huang, D., Zhang, T., Li, Y., Zhao, L., Li, J., Fang, Z., Xia, C., Li, L., He, X.: Dexterous hand manipulation via efficient imitation-bootstrapped online reinforcement learning. arXiv preprint arXiv:2503.04014 (2025)
9. Jiang, Y., Yu, M., Zhu, X., Tomizuka, M., Li, X.: 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). pp. 5260–5266. IEEE (2024)
10. Jin, W.: Complementarity-free multi-contact modeling and optimization for dexterous manipulation. arXiv preprint arXiv:2408.07855 (2024)
11. Li, G., Wang, R., Xu, P., Ye, Q., Chen, J.: The developments and challenges towards dexterous and embodied robotic manipulation: A survey. arXiv preprint arXiv:2507.11840 (2025)
12. Li, H., Zhao, Q., Xu, H., Jiang, X., Ben, Q., Jia, F., Zhao, H., Xu, L., Zeng, J., Wang, H., et al.: Teleopbench: A simulator-centric benchmark for dual-arm dexterous teleoperation. arXiv preprint arXiv:2505.12748 (2025)
13. Li, K., Li, P., Liu, T., Li, Y., Huang, S.: Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6991–7003 (2025)
14. Lin, X., Yao, K., Xu, L., Wang, X., Li, X., Wang, Y., Li, M.: Dexflow: A unified approach for dexterous hand pose retargeting and interaction. arXiv preprint arXiv:2505.01083 (2025)
15. Liu, Q., Cui, Y., Sun, Z., Li, G., Chen, J., Ye, Q.: Vtdexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In: The Thirteenth International Conference on Learning Representations
16. Liu, X., Lyu, K., Zhang, J., Du, T., Yi, L.: Quasisim: Parameterized quasi-physical simulators for dexterous manipulations transfer. arXiv preprint arXiv:2404.07988 (2024)
17. Lum, T.G.W., Lee, O.Y., Liu, C.K., Bohg, J.: Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration. arXiv preprint arXiv:2504.12609 (2025)
18. Luo, H., Feng, Y., Zhang, W., Zheng, S., Wang, Y., Yuan, H., Liu, J., Xu, C., Jin, Q., Lu, Z.: Being-h0: vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597 (2025)
19. Luo, J., Xu, C., Wu, J., Levine, S.: Precise and dexterous robotic manipulation via human-inthe-loop reinforcement learning. arXiv preprint arXiv:2410.21845 **2**(3) (2024)

20. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021)
21. Manawadu, U.A., Keitaro, N.: Dexterous manipulation based on object recognition and accurate pose estimation using rgb-d data. Sensors **24**(21), 6823 (2024)
22. Mandi, Z., Hou, Y., Fox, D., Narang, Y., Mandlekar, A., Song, S.: Dexmachina: Functional retargeting for bimanual dexterous manipulation. arXiv preprint arXiv:2505.24853 (2025)
23. Park, S., Lee, S., Choi, M., Lee, J., Kim, J., Kim, J., Joo, H.: Learning to transfer human hand skills for robot manipulations. arXiv preprint arXiv:2501.04169 (2025)
24. Patel, A., Shield, S.L., Kazi, S., Johnson, A.M., Biegler, L.T.: Contact-implicit trajectory optimization using orthogonal collocation. IEEE Robotics and Automation Letters **4**(2), 2242–2249 (2019)
25. Qin, Y., Wu, Y.H., Liu, S., Jiang, H., Yang, R., Fu, Y., Wang, X.: Dexmv: Imitation learning for dexterous manipulation from human videos. In: European Conference on Computer Vision. pp. 570–587. Springer (2022)
26. Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., Levine, S.: Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. arXiv preprint arXiv:1709.10087 (2017)
27. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
28. Shaw, K., Bahl, S., Sivakumar, A., Kannan, A., Pathak, D.: Learning dexterity from human hand motion in internet videos. The International Journal of Robotics Research **43**(4), 513–532 (2024)
29. Song, X., Li, Y., Zhang, Y., Liu, Y., Jiang, L.: An overview of learning-based dexterous grasping: recent advances and future directions. Artificial Intelligence Review **58**(10), 1–44 (2025)
30. Suh, H., Pang, T., Zhao, T., Tedrake, R.: Dexterous contact-rich manipulation via the contact trust region. arXiv preprint arXiv:2505.02291 (2025)
31. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: European conference on computer vision. pp. 581–600. Springer (2020)
32. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17868–17879 (2024)
33. Wu, Y.H., Wang, J., Wang, X.: Learning generalizable dexterous manipulation from human grasp affordance. In: Conference on Robot Learning. pp. 618–629. PMLR (2023)
34. Xin, C., Yu, M., Jiang, Y., Zhang, Z., Li, X.: Analyzing key objectives in human-to-robot retargeting for dexterous manipulation. arXiv preprint arXiv:2506.09384 (2025)
35. Yan, D., Wang, P., Zhang, T., Wang, X., Wang, Y., Wang, Z., Luo, Y., Huang, Y., Deng, G.: Casiahand: Design and evaluation of a 15-dof tendon-driven anthropomorphic robotic hand. IEEE Robotics and Automation Letters (2025)
36. Yang, F., Power, T., Marinovic, S.A., Iba, S., Zarrin, R.S., Berenson, D.: Multi-finger manipulation via trajectory optimization with differentiable rolling and geometric constraints. IEEE Robotics and Automation Letters (2025)
37. Yang, F., Zheng, S., Zhao, H., Zhan, Y., Li, X., Zhu, Y., Tang, C.Z.M., Wang, J.: Tracevision: Trajectory-aware vision-language model for human-like spatial understanding. arXiv preprint arXiv:2602.19768 (2026)

38. Yang, F., Zhu, Y., Li, X., Zhan, Y., Zhao, H., Zheng, S., Wang, Y., Tang, M., Wang, J.: Focus: Unified vision-language modeling for interactive editing driven by referential segmentation. Advances in Neural Information Processing Systems **38**, 33408–33437 (2026)
39. Ye, J., Wang, K., Yuan, C., Yang, R., Li, Y., Zhu, J., Qin, Y., Zou, X., Wang, X.: Dex1b: Learning with 1b demonstrations for dexterous manipulation. arXiv preprint arXiv:2506.17198 (2025)
40. Ye, S., Jang, J., Jeon, B., Joo, S., Yang, J., Peng, B., Mandlekar, A., Tan, R., Chao, Y.W., Lin, B.Y., et al.: Latent action pretraining from videos. arXiv preprint arXiv:2410.11758 (2024)
41. Yin, Z.H., Huang, B., Qin, Y., Chen, Q., Wang, X.: Rotating without seeing: Towards in-hand dexterity through touch. arXiv preprint arXiv:2303.10880 (2023)
42. Yuan, T., Fan, K., Ye, W., Guo, K., Guan, B., Liu, K., Kong, C., Li, J., Zhao, C., Wang, J.: Ideavatar: Identity-preserving avatar generation with controllable emotions. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1801–1805. IEEE (2026)
43. Yuan, Z., Wei, T., Gu, L., Hua, P., Liang, T., Chen, Y., Xu, H.: Hermes: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation. arXiv preprint arXiv:2508.20085 (2025)
44. Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., Lu, C.: Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 445–456 (2024)
45. Zhang, M., Jha, D.K., Raghunathan, A.U., Hauser, K.: Simultaneous trajectory optimization and contact selection for contact-rich manipulation with high-fidelity geometry. IEEE Transactions on Robotics (2025)
46. Zhao, J., Adelson, E.H.: Gelsight svelte: A human finger-shaped single-camera tactile robot finger with large sensing coverage and proprioceptive sensing. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8979–8984. IEEE (2023)
47. Zhong, Y.D., Dey, B., Chakraborty, A.: Extending lagrangian and hamiltonian neural networks with differentiable contact models. Advances in Neural Information Processing Systems **34**, 21910–21922 (2021)
48. Zhou, H., Wang, R., Tai, Y., Deng, Y., Liu, G., Jia, K.: You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations. arXiv preprint arXiv:2501.14208 (2025)
49. Zhu, H., Gupta, A., Rajeswaran, A., Levine, S., Kumar, V.: Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 3651–3657. IEEE (2019)