Learning Diverse Bimanual Dexterous Manipulation Skills from Human Demonstrations

Bohan Zhou[1], **Haoqi Yuan**[1], **Yuhui Fu**[1], **Zongqing Lu**[1,2*†]

[1]School of Computer Science, Peking University, China
[2]BeingBeyond
{zhoubh@stu., yhq@, fuyh@stu., zongqing.lu@}pku.edu.cn,

Abstract

Bimanual dexterous manipulation is a critical yet underexplored area in robotics. Its high-dimensional action space and inherent task complexity present significant challenges for policy learning, and the limited task diversity in existing benchmarks hinders general-purpose skill development. Existing approaches largely depend on reinforcement learning, often constrained by intricately designed reward functions tailored to a narrow set of tasks. In this work, we present a novel approach for efficiently learning diverse bimanual dexterous skills from abundant human demonstrations. Specifically, we introduce **BiDexHD**, a framework that unifies task construction from existing bimanual datasets and employs teacher-student policy learning to address all tasks. The teacher learns state-based policies using a general two-stage reward function across tasks with shared behaviors, while the student distills the learned multi-task policies into a vision-based policy. With BiDexHD, scalable learning of numerous bimanual dexterous skills from auto-constructed tasks becomes feasible, offering promising advances toward universal bimanual dexterous manipulation. Experiments on TACO tool-using dataset spanning 141 tasks across 6 categories demonstrate a task fulfillment rate of **74.59%** on trained tasks and **51.07%** on unseen tasks. We further transfer BiDexHD to 11 ARCTIC collaborative tasks and achieve an average of **80.49%** task fulfillment rate on trained tasks and **65.99%** on unseen task. All empirical results demonstrate the effectiveness and competitive zero-shot generalization capabilities of BiDexHD.

Introduction

Bimanual manipulation is crucial. Humans use both hands to do manipulations like using scissors or tying shoelaces (Zhou et al. 2025). The ability to manipulate objects with two hands is fundamental for everyday tasks because, with both hands, we can not only do "symmetry" collaborative tasks like carrying a heavy box but also "asymmetry" tasks (Liu et al. 2024a) like twisting a bottle cap, which means one auxiliary hand focuses on stabilizing objects and the other acts as an operator.

With the rapid development of embodied artificial intelligence, robotic bimanual dexterous manipulation is getting more and more important in manufacturing, healthcare, agriculture, and tertiary industry (Zhang et al. 2024b). Despite significant, achieving proficient bimanual manipulation remains a substantial challenge because it severely struggles with high-dimensional action spaces. While a line of previous work (Grannen et al. 2023; Yu et al. 2024; Liu et al. 2024a) primarily focuses on bimanual grippers, bimanual manipulation with dexterous hands is still largely unexplored. Previous attempts to solve bimanual dexterous tasks are mainly based on reinforcement learning (RL) (Lin et al. 2024; Huang et al. 2023; Zhang et al. 2024a). However, they require intricate reward designs tailored to specific manually-designed tasks. Therefore, these approaches lack scalability and generalizability to a broader range of tasks. Recent research (Sindhupathiraja et al. 2024; Fu et al. 2024; He et al. 2024) has advanced bimanual dexterous manipulation through teleoperation. Nevertheless, human intervention is inevitable. We would ask a question: ***Can we learn diverse bimanual dexterous manipulation skills in a unified and scalable way?***

Our answer is to leverage human demonstrations, which are easier to collect via haptic gloves or MoCap devices than robotic rollouts, and offer more compliant, human-aligned behaviors, inspired by recent advances in learning from demonstration (Zhou et al. 2023, 2024). In this paper, we propose a novel approach to learn diverse bimanual dexterous manipulation skills from human demonstrations. Upon this setting, we propose **BiDexHD**, a framework to automatically turn a human bimanual manipulation dataset into a series of tasks in simulation and conduct effective teacher-student learning. The majority of previous bimanual studies primarily focus on existing benchmarks or a limited range of tasks. For RL-based methods (Lin et al. 2024; Huang et al. 2023; Zhang et al. 2024a), they tailor specific reward function to specific tasks. For IL-based methods (Wang et al. 2024a; Cheng et al. 2024), it is inevitable to collect a bulk of data for learning specific tasks (typically around 50 trajectories each task). BiDexHD does not depend on manually-designed tasks or pre-defined tasks in existing benchmarks, and instead, unifies task construction from any accessble bimanual manipulation trajectory, which makes it scalable. Furthermore, BiDexHD gets rid of task-specific reward engineering, and instead, designs a general reward function for all automatically constructed object-centric tasks. In a word, BiDexHD is such a unified and scalable framework that breaks the bottleneck

*Corresponding author.
†Done at the Beijing Academy of Artificial Intelligence.

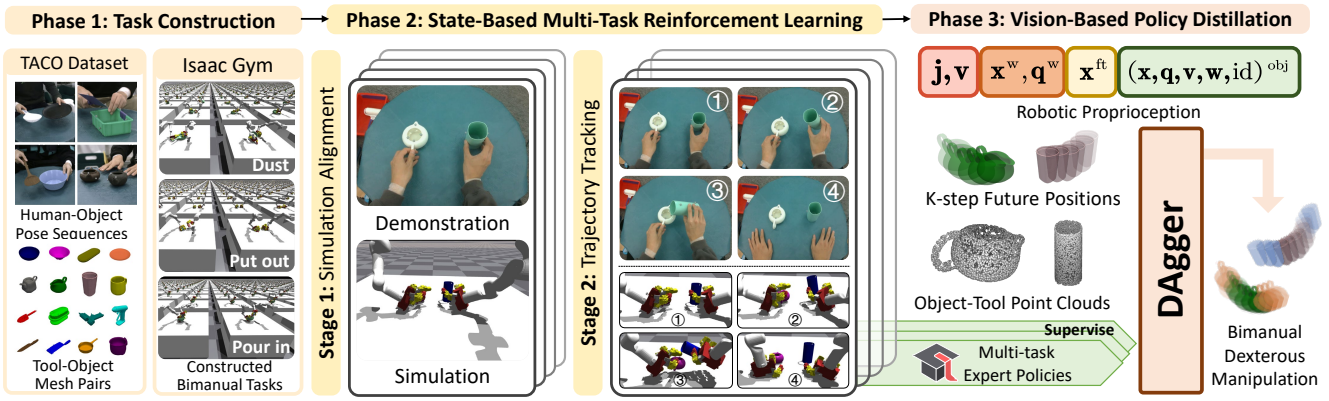

Figure 1: The three-phase framework, BiDexHD, unifies constructing and solving tasks from human bimanual datasets instead of existing benchmarks. In phase one, BiDexHD constructs each bimanual task from a human demonstration. In phase two, BiDexHD learns diverse state-based policies from a generally designed two-stage reward function via multi-task reinforcement learning. A group of learned policies are then distilled into a vision-based policy for inference in phase three.

of limited tasks and label-intensive manual designs, which is meaningful to the further development of general-purpose bimanual dexterous manipulation. Though promising, several challenges must be addressed to fully realize this. It is essential to figure out how to accurately mimic fine-grained bimanual behaviors from human demonstrations and avoid collisions and disturbances while encouraging smooth trajectories and synchronous collaboration. To address this, we design a general two-stage reward function to assign curricula for RL training, which is empirically proven effective in bimanual tool-using tasks constrcuted from TACO (Liu et al. 2024b) and collaborative tasks constrcuted from ARCTIC (Fan et al. 2023).

To sum up, three key contributions can be summarized:

- We propose **a unified and scalable framework BiDexHD**, which unifies automatic task construction from HOI datasets and teacher-student policy learning to learn diverse vision-based bimanual dexterous skills from human demonstrations.
- To avoid task-specific reward engineering, a **two-stage reward function** is generally designed to guide multi-task state-based policy learning on all object-centric tasks.
- We evaluate BiDexHD across **141 auto-constructed tool-using tasks over 6 categories from TACO** (Liu et al. 2024b) dataset and **11 collaborative tasks from ARCTIC** (Fan et al. 2023) dataset. The results demonstrate BiDexHD's zero-shot capabilities and scalability.

Related Work

Bimanual Dexterous Manipulation. In recent years, robotics research highlights dexterous manipulation for its human-like flexibility. Dexterous hands enable tasks including in-hand manipulation (Arunachalam et al. 2023; Yin et al. 2023; Qi et al. 2023; Chen, Xu, and Agrawal 2022), grasping (Xu et al. 2023; Wan et al. 2023; Yuan et al. 2024; Huang et al. 2024), and deformable object manipulation (Bai, Yu, and Liu 2016; Li et al. 2023; Hou, Sahari, and How 2019). However, most existing studies focus on single hand, concealing the potential of bimanual dexterity. Most related work is PGDM (Dasari, Gupta, and Kumar 2023) and we compare BiDexHD with it in Appendix. Existing bimanual research explores varied approaches: RL for specific tasks (e.g., Lin et al. (2024); Huang et al. (2023); Zhang et al. (2024a)), LLM-based systems Gbagbe et al. (2024), physics-aware learning Wang et al. (2024b), and keypoints-based imitation Gao et al. (2024). Unlike these, **BiDexHD** provides a general framework leveraging unified reward functions for state-based RL training.

Learning Dexterity From Human Demonstrations. Learning from human demonstrations provides efficient, data-driven solutions for robot learning (Jia et al. 2024; Odesanmi, Wang, and Mai 2023; Yuan et al. 2025; Luo et al. 2025), avoiding RL's high-DoF and reward engineering challenges (Smith et al. 2019; Schmeckpeper et al. 2020; Shao et al. 2021). Prior works (Arunachalam et al. 2023; Mandikal and Grauman 2021; Sivakumar, Shaw, and Pathak 2022; Qin et al. 2022; Mandikal and Grauman 2022; Liu et al. 2023; Shaw, Bahl, and Pathak 2023; Chen et al. 2024) focus on single-hand tasks (e.g., in-hand manipulation (Arunachalam et al. 2023)) or video-conditioned teleoperation (Sivakumar, Shaw, and Pathak 2022). Recent bimanual datasets (Zhan et al. 2024; Liu et al. 2024b; Fan et al. 2023; Razali and Demiris 2023) offer rich dual-hand poses and hand-object interactions, enabling automated task definition. In this work, we tackle general bimanual skill learning via tasks auto-constructed from these demonstrations.

Preliminaries

Task Formulation. We formulate each bimanual manipulation task as a decentralized partially observable Markov decision process (Dec-POMDP). The Dec-POMDP can be represented by the tuple $Z = (\mathcal{N}, \mathcal{M}, S, \boldsymbol{O}, \boldsymbol{A}, P, R, \rho, \gamma)$. Dual hands with arms are separated as $\mathcal{N}$ agents, which is represented by set $\mathcal{M}$. The observation space $\boldsymbol{O}$ contains robot proprioception and object information, which are ini-

tialized at $s_0 \in S$ according to the initial state distribution $\rho(s_0)$. At each time step t, the state is represented by s_t, and the i-th agent receives an observation $o_t^i \in \boldsymbol{O}$ based on s_t. Subsequently, the policy of the i-th agent, $\pi_i \in \boldsymbol{\Pi}$, takes o_t^i as input and outputs an action $a_t^i \in A^i$. The joint action of all agents is denoted by $\boldsymbol{a}_t \in \boldsymbol{A}$, where $\boldsymbol{A} = A^1 \times A^2 \times \cdots A^{\mathcal{N}}$. The state transits to the next state according to the transition function $s_{t+1} \sim P(s_{t+1}|s_t, \boldsymbol{a}_t)$. After this, the i-th agent receives a reward r_t^i based on the reward function $R(s_t, \boldsymbol{a}_t)$. The objective is to find the optimal policy $\boldsymbol{\pi}$ that maximizes the expected sum of rewards $\mathbb{E}_{\boldsymbol{\pi}}[\sum_{t=0}^{T-1} \gamma^t \sum_{i=1}^{\mathcal{N}} r_t^i]$ over an episode with T time steps, where γ is the discount factor.

Dataset Preparation. A human bimanual manipulation dataset consists of M trajectories $\mathcal{D} = \{\tau^1, \tau^2, \ldots, \tau^M\}$, each of which describes a human using a tool with his right hand to manipulate a target object with his left hand. The behavior of each trajectory can be recapped with a triplet (action, tool, object). Any triplet belongs to a union $\mathcal{U} = \mathcal{V} \times \Omega \times \Omega$, where Ω denotes the set of all objects and tools, and $\mathcal{V}$ denotes the set of all human actions. According to different behaviors depicted in $\mathcal{V}$, we can split all the tasks into $|\mathcal{V}|$ categories. Each trajectory $\tau^i = \{\mathbf{h}^{\text{tool}}, \mathbf{h}^{\text{object}}, \hat{\mathbf{x}}_t^{\text{tool}}, \hat{\mathbf{q}}_t^{\text{tool}}, \hat{\mathbf{x}}_t^{\text{object}}, \hat{\mathbf{q}}_t^{\text{object}}, \Theta_t^{\text{left}}, \Theta_t^{\text{right}}\}_{t:1..N}^i$ involves a pair of meshes of the tool and object from a object mesh set $\mathbf{h}^{\text{tool}}, \mathbf{h}^{\text{object}} \in \mathcal{H}$, N-step position $\mathbf{x} \in \mathbb{R}^3$ and orientation $\mathbf{q} \in \mathbb{R}^4$ sequence of the tool and the object, and the pose sequence of hands described in MANO (Romero, Tzionas, and Black 2017) parameters Θ.

Teacher-Student Learning. Directly learning a multi-task vision-based dexterous policy is extremely challenging (Chen, Xu, and Agrawal 2022; Chen et al. 2023). Teacher-student learning (Wan et al. 2023) is a more progressive and scalable framework. In the teacher learning phase, multiple state-based policies is first trained via reinforcement learning, leveraging privileged information to solve multiple similar tasks. In the student learning phase, a vision-based student policy is distilled from a bunch of teacher policies. A key difference in observation is that the teacher's view includes precise object state details, while the student relies on point clouds of P sampled points from the object's surface. This makes the student policy suitable for real-world deployment, given real-time point clouds from a multi-view RGB-D camera system.

Learning Bimanual Dexterity From Human Demonstrations

Overview

As illustrated in Figure 1, we propose a scalable three-phase framework. In the first phase, we parallelize the construction of Dec-POMDP bimanual tasks from a human bimanual manipulation dataset within IsaacGym (Makoviychuk et al. 2021). After task initialization, the subsequent two phases adopt a teacher-student policy learning framework. Following the approach of Chen, Xu, and Agrawal (2022); Chen et al. (2023); Wan et al. (2023), we utilize Independent Proximal Policy Optimization (IPPO) (De Witt et al. 2020) during the second phase to train state-based teacher policies in parallel independently. Each expert focuses on a subset of tasks that require similar behaviors. In the final phase, the teacher policies are distilled into a vision-based student policy.

Task Construction From Bimanual Dataset

Data Preprocessing. We extract the wrist and fingertip pose of dual hands at each timestep $\{V_t^{\text{side}}, J_t^{\text{side}}\} = \text{MANO}(\Theta_t^{\text{side}})$, side $\in$ {left, right} with MANO (Romero, Tzionas, and Black 2017) parameters $\Theta = \{\alpha, \beta, \hat{\mathbf{x}}^{\text{w}}\}$, where $\alpha \in \mathbb{R}^{48}, \beta \in \mathbb{R}^{10}$, and $\hat{\mathbf{x}}^{\text{w}} \in \mathbb{R}^3$ represent hand pose, hand shape parameters, and wrist position respectively. $V \in \mathbb{R}^{778\times3}$ and $J \in \mathbb{R}^{21\times3}$ represent vertices and joints on a hand respectively. The quaternion of the wrist $\hat{\mathbf{q}}^{\text{w}} \in \mathbb{R}^4$ is translated from axis-angle $\beta_{0:3}$. Given that single LEAP Hand (Shaw, Agarwal, and Pathak 2023) has only four fingers, we can easily filter the corresponding positions of these $m = 4$ fingers in J, denoting them as $\mathbf{x}^{\text{ft}} \in \mathbb{R}^{m\times3}$. In the following sections, τ^i is denoted as:

$$\tau^i = \{(\hat{\mathbf{x}}, \hat{\mathbf{q}})_t^{\{\text{tool,object,lw,rw}\}}, \hat{\mathbf{x}}_t^{\text{lft}}, \hat{\mathbf{x}}_t^{\text{rft}}\}_{t:1..N}^i. \quad (1)$$

Simulation Initialization. After data preprocessing, bimanual manipulation tasks $\Gamma = \{\mathcal{T}^1, ..., \mathcal{T}^M\}$ can be constructed in Issac Gym in parallel. For each task $\mathcal{T}^i$, the mesh of a tool $\mathbf{h}^{\text{tool}}$ and a target object $\mathbf{h}^{\text{object}}$, along with two arms with hands are initialized with a fixed state vector:

$$o_0^{\text{side}} = \{(\mathbf{j}, \mathbf{v})^{\text{side}}, (\mathbf{x}, \mathbf{q})^{\text{side,w}}, \mathbf{x}^{\text{side,ft}}, (\mathbf{x}, \mathbf{q}, \mathbf{v}, \mathbf{w}, \text{id})^{\text{obj}}\}_0^{\text{side}}$$
$$\text{where side, obj} \in \{(\text{left, object}), (\text{right, tool})\}.$$

The robot proprioception includes arm-hand joint angles and velocities, wrist poses, and fingertip positions, and object information includes object positions, orientations, linear and angular velocities, and a unique object identifier for multi-task learning. For all tasks, $(\mathbf{j}, \mathbf{v})_0$ are all reset to zero. The initial states of wrist and fingertips are calculated with forward kinematics accordingly. Except identifiers, the initial observations for all tools and target objects keep unchanged. It is worth noting that we assume the robot to be right-handed by default, *i.e.*, the left hand handles the target object and the right hand handles the tool. For brevity, the repeated notation side, obj $\in$ {(left, object),(right, tool)} is omitted in subsequent sections.

To ensure the feasibility of each task, after initialization, hand joint angles are optimized from human hand motions via AnyTeleop Qin et al. (2023) and arm joint angles are calculated via inverse kinematics (IK) based on the robot's palm base pose. By replaying all object-hand trajectories in the simulator, we can easily identify and remove invalid tasks to build up a complete task set Γ.

Multi-Task State-Based Policy Learning

In the second phase, we focus on learning a multi-task state-based policy for tasks that require similar behaviors. We can broadly divide these tasks into two stages: First, aligning the simulation state with initial τ_0^i of a trajectory, and second, following each step of the trajectory. During the alignment stage, both hands should prioritize approaching their objects. The left hand learns to grasp or stabilize the target object, while the right hand learns to grasp the tool. Once simulation alignment is achieved, both

hands are expected to maintain their hold and follow the pre-defined trajectory derived from the human demonstration dataset to perform the manipulations in sync. The pipeline is illustrated in Figure 2. We initialize objects and robots at stage zero, finish simulation alignment at stage one, and conduct trajectory tracking at stage two via IPPO to learn state-based policies $\pi_\theta^{\text{side}}(\mathbf{a}_t^{\text{side}}|o_t^{\text{side}}, \mathbf{a}_{t-1}^{\text{side}})$ conditioning on the current observation $o_t^{\text{side}} = \{(\mathbf{j}, \mathbf{v})^{\text{side}}, (\mathbf{x}, \mathbf{q})^{\text{side,w}}, \mathbf{x}^{\text{side,ft}}, (\mathbf{x}, \mathbf{q}, \mathbf{v}, \mathbf{w}, \text{id})^{\text{obj}}\}_t^{\text{side}}$ and previously executed action $\mathbf{a}_{t-1}^{\text{side}}$ for dual hands.

Stage 1: Simulation Alignment. The central goal of stage one is to align the state of simulation to the first step in a trajectory by moving the tool and target object from the fixed initial pose to τ_0, which serves as an essential yet challenging prerequisite for subsequent trajectory tracking in stage two. Through experiments in later sections, we find that it is not feasible to directly acquire dynamic skills from static poses through imitation. Instead, we adopt reinforcement learning to develop skills like grasping, twisting and pushing. Some previous work (Luo et al. 2024; Xu et al. 2023) on grasping prefers introducing additional pre-grasp poses by estimating grasping pose upon manipulated objects. We adopt a simpler but more generalizable approach by learning skills directly from the object poses provided in the dataset. Specifically, we anchor the first timestep in the dataset as the reference timestep to establish a tool-object reference pose pair for each manipulation task. Stage one is considered complete once both the tool and the object reach the specified pose for a sustained u-step duration. See Appendix for examples. Rewards are carefully designed to encourage the object to be lifted above the table in reference to the filtered reference poses. The total reward consists of an approaching reward, a lifting reward, and a bonus reward.

The **approaching reward**, r_{appro}, encourages both dexterous hands to approach and remain close to the object. In other words, the goal is to minimize the distance between the robot's palm, fingertips, and the grasp center. Since functional grasping is critical for tool using, we do not simply select the geometric center of the object. Instead, we precompute the grasping center $\hat{\mathbf{x}}_{\text{gc}}$ for each tool and object based on the dataset. Specifically, for each task, we use the human-demonstrated wrist and fingertip positions at the reference timestep–$\hat{\mathbf{x}}_0^{\text{lw}}, \hat{\mathbf{x}}_0^{\text{rw}}, \hat{\mathbf{x}}_0^{\text{lft}}, \hat{\mathbf{x}}_0^{\text{rft}}$–as anchor points. We then uniformly sample 1024 points from the surface of the object mesh $\mathbf{h}^{\text{tool}}, \mathbf{h}^{\text{object}}$ to form a representative point set $\mathcal{P}$ and compute the average grasp center based on the top $L = 50$ nearest points. r_{appro} penalizes the distance between the wrist, fingertips, and the grasp center, and is defined as

$$r_{\text{appro}}^{\text{side}} = -\|\mathbf{x}_t^{\text{side,w}} - \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}}\|_2 - w_r \sum^m \|\mathbf{x}_t^{\text{side,ft}} - \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}}\|_2$$
$$\text{where } \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}} = \frac{1}{L} \sum \text{NN}\left(\mathcal{P}, L, \frac{\hat{\mathbf{x}}_0^{\text{side,w}} + \sum^m \hat{\mathbf{x}}_0^{\text{side,ft}}}{m+1}\right). \tag{2}$$

The **lifting reward** r_{lift} encourages holding objects tightly in hands and lifting to desired reference poses. The robots receive a lifting reward r_{lift} composed of a non-negative linear position reward and a negative quaternion distance reward if the lifting conditions are satisfied. $\mathbf{x}_0^{\text{object}}$ and $\mathbf{x}_0^{\text{tool}}$ represent the initial positions of the target object and tool in the simulator. $\hat{\mathbf{x}}_0$ denotes the first reference position in a human demonstration.

$$r_{\text{pos}}^{\text{side}} = \max\left(1 - \frac{\|\mathbf{x}_t^{\text{obj}} - \hat{\mathbf{x}}_0^{\text{obj}}\|_2}{\|\mathbf{x}_0^{\text{obj}} - \hat{\mathbf{x}}_0^{\text{obj}}\|_2}, 0\right),$$
$$r_{\text{quat}}^{\text{side}} = -\mathbb{D}_{\text{quat}}\left(\mathbf{q}_t^{\text{obj}}, \hat{\mathbf{q}}_0^{\text{obj}}\right),$$
$$r_{\text{lift}}^{\text{side}} = \left(r_{\text{pos}}^{\text{side}} + w_q r_{\text{quat}}^{\text{side}}\right) \cdot \mathbb{I}\left(\|\mathbf{x}_t^{\text{side,w}} - \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}}\|_2 \le \lambda_{\text{w}}\right) \cdot \mathbb{I}\Big(\sum^m \|\mathbf{x}_t^{\text{side,ft}} - \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}}\|_2 \le \lambda_{\text{ft}}\Big).$$

The **bonus reward** r_{bonus} incentivizes the target object or the tool to reach and finally stay at their reference poses, which lays a foundation for the second manipulation stage. r_{bonus} becomes positive only when the distance between an object's current position and its reference position becomes lower than $\varepsilon_{\text{succ}}$. Stage one is considered successful only if both $r_{\text{bonus}}^{\text{left}}$ and $r_{\text{bonus}}^{\text{right}}$ are positive for at least u consecutive steps. Thus, the bonus reward r_{bonus} is defined as

$$r_{\text{bonus}}^{\text{side}} = \begin{cases} \frac{1}{1+\|\mathbf{x}_t^{\text{obj}} - \hat{\mathbf{x}}_0^{\text{obj}}\|_2} & \text{if } \mathbb{I}\left(\|\mathbf{x}_t^{\text{obj}} - \hat{\mathbf{x}}_0^{\text{obj}}\|_2 \le \varepsilon_{\text{succ}}\right) \\ 0 & \text{otherwise.} \end{cases}$$

The total alignment reward is the linear weighted sum of the three components.

$$r_{\text{align}}^{\text{side}} = w_1 r_{\text{appro}}^{\text{side}} + w_2 r_{\text{lift}}^{\text{side}} + w_3 r_{\text{bonus}}^{\text{side}}$$

Stage 2: Trajectory Tracking. Once stage one is completed, the left hand is securely holding the target object, and the right hand keeps grasping the tool at its desired reference pose. The next step is to maintain the grasp and follow a trajectory to perform the manipulation. To achieve this, we design a more fine-grained exponential reward, r_{track}, which encourages the dexterous hands to precisely track the desired positions at each timestep in a trajectory starting from the reference timestep. Assuming that human hands are more flexible than robotic hands, we introduce a constant tracking frequency f, where f simulation steps correspond to one step in the dataset. Let $\hat{\mathbf{x}}_i^{\text{obj}}$ represent the position of a object at i-th step in a l-step human-demonstrated trajectory and $\mathbf{x}_{t_i}^{\text{obj}}$ represent the object's position at the corresponding simulation step in IsaacGym. We have $i = \lceil t_i/f \rceil \in [0, l)$, and the tracking reward is defined as:

$$r_{\text{track}}^{\text{side}} = \begin{cases} \exp\left(-w_t \|\mathbf{x}_{t_i}^{\text{obj}} - \hat{\mathbf{x}}_i^{\text{obj}}\|_2\right) & \text{if stage 1 succeeds} \\ 0 & \text{otherwise.} \end{cases}$$

We adopt IPPO to learn a unified policy from the combination of all rewards for the two stages $r_{\text{total}}^{\text{side}} = r_{\text{align}}^{\text{side}} + w_4 r_{\text{track}}^{\text{side}}$. r_{total} unifies two stages of bimanual dexterous manipulation, enabling scaling up to multi-task policy learning for a wide range of constructed bimanual tasks.

Vision-Based Policy Distillation

We employ DAgger (Ross, Gordon, and Bagnell 2011), an on-policy imitation learning algorithm, to develop a vision-based policy for each task category $\nu \in \mathcal{V}$, under the supervision of a group of state-based teacher poli-

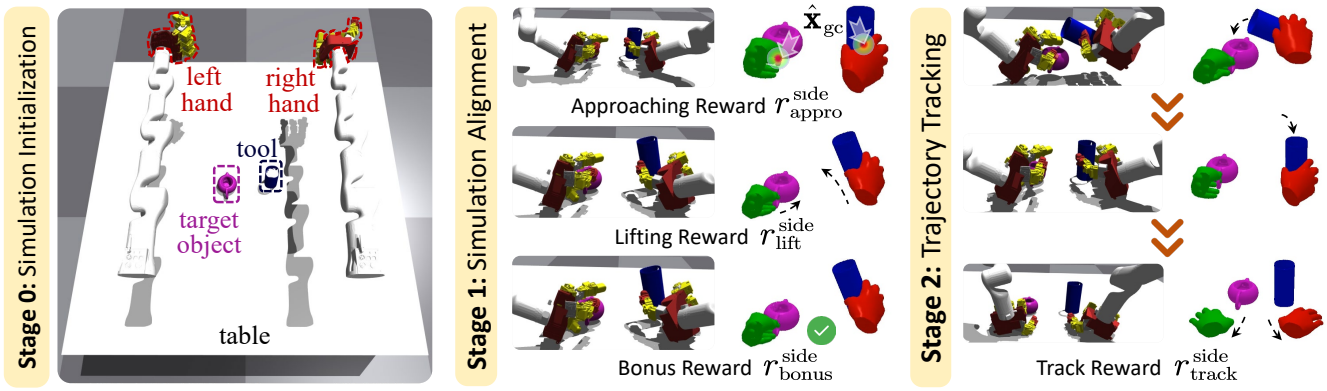

Figure 2: General two-stage teacher learning. For each task $\mathcal{T}^i$, all zero poses are initialized at stage zero. Both the tool and object are initialized to poses sampled from a fixed Gaussian distribution centered at a fixed value with added small noise. At stage one, approaching reward r_{appro} encourages both hands to get close to their grasping centers $\hat{\mathbf{x}}_{\text{gc}}$, and lifting reward r_{lift} along with extra bonus r_{bonus} incentivizes moving both objects to thier reference poses respectively. After simulation alignment, dual hands will manipulate objects under the guidance of tracking reward r_{track}.

cies. To enhance generalization capabilities for new objects or unseen tasks, we propose transforming the student policy into a trajectory-conditioned in-context policy, denoted as $\pi^{\text{side}}_{\phi}(\mathbf{a}^{\text{side}}_t|\mathbf{o}^{\text{side}}_t, \mathbf{p}^{\text{side}}_t, \mathbf{a}^{\text{side}}_{t-1})$, where $\mathbf{o}_t = \{(\mathbf{j},\mathbf{v})^{\text{side}}, (\mathbf{x},\mathbf{q})^{\text{side,w}}, \mathbf{x}^{\text{side,ft}}, \text{pc}^{\text{obj}}\}_t$, K-step future pose $\mathbf{p}^{\text{side}}_t \in \mathbb{R}^{K\times 3}$, and $\text{pc}^{\text{obj}}_t \in \mathbb{R}^{P\times 3}$. Specifically, to get point clouds $\text{pc}^{\text{tool}}_t$ and $\text{pc}^{\text{object}}_t$, we pre-sample 4096 points from the surface of $\mathbf{h}^{\text{tool}}$ and $\mathbf{h}^{\text{object}}$ for each task during initialization. At each timestep t, a subset of points are sampled from the pre-sampled point clouds, transformed according to current object pose and added with Gaussian noise for robustness. Besides, during DAgger distillation, we augment traditional vision-based policy $\pi^{\text{side}}_{\phi}(\mathbf{a}^{\text{side}}_t|\mathbf{o}^{\text{side}}_t, \mathbf{a}^{\text{side}}_{t-1})$ with next K positions along the object's trajectory as additional inputs. This design allows the learned policy to utilize more information about the motion of objects, such as movement direction and speed in the near future, facilitating zero-shot transfer to unfamiliar tasks or objects. Notably, we can easily mask this additional input by setting $K = 0$. We further investigate the influence of K future steps in Section . Refer to Appendix for algorithm and implementation details.

Experiments

Dataset. We mainly evaluate the effectiveness of BiDexHD on the TACO (Liu et al. 2024b) dataset, a large-scale bimanual manipulation dataset that encompasses diverse human demonstrations using tools to manipulate target objects in real-world scenarios. BiDextHD converts 6 categories $\mathcal{V} = \{\text{Dust, Empty, Pour in some, Put out, Skim off, Smear}\}$ of total 141 human demonstrations in the TACO dataset to Dec-POMDP tasks (See Appendix for task examples). Task diversity and abundance make BiDexHD easy to scale up. All tasks can be separated into 16 semantic groups, each of which gathers a number of similar demonstrations with the same action, the same tool-object category but different tool and object instances. BiDextHD constructs a task from single demonstration, and thus each semantic group correspond to a semantic subtask. We adopt teacher-student learning to train 16 semantic sub-tasks and distill teacher policies with similar skills into 6 vision-based policies for each category eventually. We split 80% tasks for training (**Train**) and the rest 20% unseen tasks for testing. Detailed descriptions of dataset split are provided in Appendix. For each task in the testing set, if the object and tool both occur in the training set it is labeled as a kind of combinational task (**Test Comb**), and otherwise it is labeled as a new task (**Test New**).

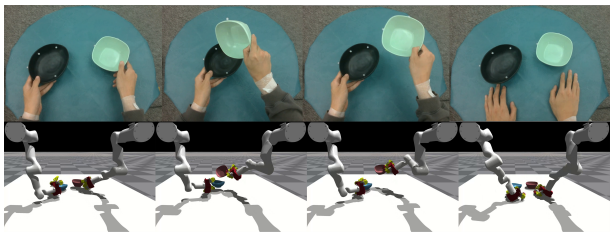

Figure 3: Visualization of sampled task: (empty, bowl, bowl).

Metrics. To measure the quality of our constructed tasks, we introduce two metrics m_1 and m_2.

- m_1: the average success rate of stage one. For n episodes, $\text{m}_1 = \frac{1}{n}\sum_{e=1}^{n}\mathbb{I}_1^e$ averages over the number of episodes that satisfys conditions $\mathbb{I}_1$ at stage one.

$$\mathbb{I}_1 : \exists 0 < t < T-u \sum_{t}^{t+u} \prod^{\{\text{tool,object}\}} \mathbb{I}\Big(\|\mathbf{x}^{\text{obj}}_t - \hat{\mathbf{x}}^{\text{obj}}_0\|_2 \le \varepsilon_{\text{succ}}\Big) \cdot \mathbb{I}\Big(\mathbb{D}_{\text{quat}}(\mathbf{q}^{\text{obj}}_t, \hat{\mathbf{q}}^{\text{obj}}_0) \le \varepsilon_{\text{succ}}\Big) = u$$

- m_2: the average tracking rate of stage two. Each task corresponds to l-step human demonstration. For each episode, calculate the proportion of steps where two objects both effectively follows their desired poses. m_2 is the average

tracking rate over n episodes.

$$\mathrm{m}_2 = \frac{1}{nl}\sum_{i=0}^{n}\sum_{}^{l-1}\prod^{\{\text{tool,object}\}} \mathbb{I}\Big(\|\mathbf{x}_{t_i}^{\text{obj}} - \mathbf{x}_i^{\text{obj}}\|_2 \leq \varepsilon_{\text{track}}\Big)\cdot \mathbb{I}\Big(\mathbb{D}_{\text{quat}}(\mathbf{q}_{t_i}^{\text{obj}}, \mathbf{q}_i^{\text{obj}}) \leq \varepsilon_{\text{track}}\Big)$$

It is important to note that m_2 serves as the primary metric for indicating task completion while m_1 is an intermediate metric for assessing task progression. Considering the choice of $\varepsilon_{\text{succ}}$ and $\varepsilon_{\text{track}}$ has a non-legligible impact for evaluation, we will discuss the sensitivity of these thresholds in Appendix. By default, we choose $\varepsilon_{\text{succ}} = \varepsilon_{\text{track}} = 0.1$.

Teacher Learning

Different RL algorithms can be incorporated into BiDexHD. We mainly compare the performance of independent PPO (**BiDexHD-IPPO**) and centralized PPO (**BiDexHD-PPO**). For BiDexHD-IPPO, two agents possess their own observations and execute their own actions. For BiDexHD-PPO, a single policy takes as input both observations and is trained to output all actions that maximize the sum of all total rewards in an episode, which transforms a Dec-POMDP task into a POMDP task.

RL Results. Table 1 (green rows) shows average performance on auto-constructed bimanual tasks. BiDexHD-IPPO achieves near-complete stage-one success and high tracking quality for seen objects (Train/Test Comb), demonstrating scalability across TACO tasks. BiDexHD-PPO underperforms due to IPPO's superior efficiency in learning robust skills via independent left/right policies with smaller observation/action spaces. Furthermore, two independent expert policies focusing solely on specific groups of target objects or tools adapt more easily to similar combinational tasks than a single policy that must attend to both. See Appendix for more detailed discussion and evaluation results. Both BiDexHD-IPPO/PPO experience a noticeable performance decline on new objects (Test New). Both methods decline on new objects (Test New) due to one-hot object label dependency in state-based training. The primary reason for this drop is that these approaches incorporate one-hot object labels in observations during state-based training, leading the policy to heavily rely on this information. We address this by removing labels during vision-policy distillation to enhance generalization.

Ablations on Teacher Learning

Alignment Stage. To demonstrate the necessity of the design of dataset-simulation alignment stage, we compare BiDexHD-IPPO with a more naive version, denoted as **(w/o stage-1)**, which retains only r_{track} in RL training at stage 2 and maintains a fixed number of free exploration steps at stage 1. The second line in the green section of Table 1 reveals a significant performance decline. We observe that only 30.5% of relatively easy tasks (See Appendix for details) achieve positive m_1 and m_2, while for the remaining tasks, the success rate of stage 1 and the tracking rate of stage 2 remain at zero, which emphasizes the importance of r_{align} during stage 1.

Functional Grasping Center. In BiDexHD, we pre-compute the grasping center $\hat{\mathbf{x}}_{\text{gc}}$ to calculate r_{appro} in Equation 2. In this section, we replace the grasping center with the object geometric center, denoted as **(w/o gc)**. The results presented in the third line of Table 1 show a decrease in m_1 and m_2, particularly on tasks involving seen objects compared to BiDexHD-IPPO. To further investigate their behavior discrepancy, we visualize their grasping poses for a typical task (dust, brush, pan) in Appendix. BiDexHD-IPPO aligns more closely with the calculated grasping centers (red points), exhibiting human-like grasping behavior. In contrast, BiDexHD-IPPO (w/o gc) with geometric centers (green points) struggles to find proper poses for using the brush or holding the pan. In fact, the geometric center of an object does not often fall within areas suitable for manipulation. These findings highlight the significance of incorporating a functional grasping center, particularly for objects that are thin, flat, or equipped with handles.

Success Bonus. The 4^{th} line in the green section of Table 1 investigates whether removing reward r_{bonus} defined in Equation affects performance. We observe a decline in m_2 on both the training set and unseen tasks involving new objects. We analyze the additional bonus in Equation effectively signals the transition between the two stages, enhancing the policy's awareness of task progression.

Student Learning

For the BiDexHD variants, several trained multi-task state-based teacher policies from one task category are distilled into a single vision-based policy, which is then tested on all tasks. We also introduce behavior cloning (BC) as our baseline, whose configuration is reported in Appendix. To directly learn bimanual skills from a dataset, we employ Dexpilot (Handa et al. 2020) to retarget human hand motions in the TACO dataset to joint angles for dexterous hands, solving IK for arm joint angles. All joint angles are collected and replayed in IsaacGym (Makoviychuk et al. 2021) to gather observations. BC learns purely from this static observation-action dataset and is ultimately tested under the same configuration as BiDexHD.

DAgger Results. The blue section of Table 1 displays the performance of the vision-based policies. Our BiDexHD-IPPO+DAgger significantly outperforms both PPO variant and BC, achieving a high task completion rate on the training set and an average $\mathrm{m}_2 = 51.07\%$ across all unseen tasks (Test Comb and Test New). This evidence indicates the scalability and competitive generalization ability of BiDexHD framework. Among unseen tasks, we observe a slight decline in m_2 for combinational tasks, while tasks involving new objects show a sharp increase in m_2. This suggests that the vision-based policy relies more on information from the point clouds, such as shape and local features, rather than specific one-hot identifiers, enabling effective zero-shot generalization. Conversely, BC performs poorly due to the loss of true dynamics in the simulation, often getting confused by unfamiliar observations and stuck in stationary states. This also reflects the challenges associated with our constructed bimanual tasks. Our framework unifies bimanual skill learning through a combination of trial-and-error and distillation,

Method	Train $\mathrm{m}_1(\%)$	Train $\mathrm{m}_2(\%)$	Test Comb $\mathrm{m}_1(\%)$	Test Comb $\mathrm{m}_2(\%)$	Test New $\mathrm{m}_1(\%)$	Test New $\mathrm{m}_2(\%)$
BiDexHD-PPO	90.55	53.88	78.74	36.99	**81.42**	**26.24**
BiDexHD-IPPO (w/o stage-1)	25.00	17.52	24.80	18.10	19.85	08.51
BiDexHD-IPPO (w/o gc)	90.53	66.39	91.47	52.11	77.03	22.63
BiDexHD-IPPO (w/o bonus)	97.67	66.65	98.01	59.76	77.96	17.52
BiDexHD-IPPO	**98.71**	**78.18**	**98.37**	**59.94**	75.48	21.34
BC	00.00	00.00	00.00	00.00	00.00	00.00
BiDexHD-PPO+DAgger	95.35	55.82	76.75	30.42	86.34	30.00
BiDexHD-IPPO+DAgger	**99.38**	**74.59**	**92.85**	**48.43**	**94.79**	**53.71**

Table 1: The average success rate of stage 1 and tracking rate of stage 2 during training and evaluation across all tasks constructed from the TACO dataset under $\varepsilon_{\text{succ}} = \varepsilon_{\text{track}} = 0.1$.

providing a robust and scalable solution to diverse challenging bimanual manipulation tasks.Detailed evaluation results are reported in Appendix . We observe that in 'Dust' and 'Empty' task categories, the task diversity is relatively ample. Therefore, the distilled policy can surpass the average level of expert policies, which proves that distilling similarity policies bring about positive promotion effects to the final unified policy.

Future Conditioned Steps. We further examine the selection of $K \in \{0, 1, 2, 5\}$ for future object positions. Specifically, when $K = 0$, the vision-based policy relies exclusively on 3D information from object point clouds and the robot's proprioception. As shown in Table 2, the performance across different values of K does not vary significantly. Even when future conditioned steps are masked $(K = 0)$, m_2 only exhibits slight declines of 2.5% on trained tasks and an average of 3.1% on all unseen tasks compared to $K = 5$. This evidence suggests that after the multi-task RL training phase, the teachers have acquired diverse and robust skills, making pure imitation sufficient for a student to achieve acceptable performance. Nonetheless, K future steps provide additional informative and fine-grained, albeit implicit, clues such as motion and intention for more precise tracking. We discuss planning future object trajectories for real-world deployment in Appendix.

Metrics (%)	K			
	0	1	2	5
Train m_1	98.01	98.81	98.71	**99.38**
Train m_2	72.09	**75.40**	75.01	74.59
Test Comb m_1	**94.36**	92.11	93.26	92.85
Test Comb m_2	46.64	**49.02**	48.60	48.43
Test New m_1	93.96	94.67	94.38	**94.79**
Test New m_2	49.27	51.00	50.39	**53.71**

Table 2: The metrics of different K future steps.

Scalability of BiDexHD

To demonstrate BiDexHD is scalable and transferable in heterogeneous bimanual tasks, we extend our BiDexHD framework to another bimanual dataset **ARCTIC** (Fan et al. 2023), which mainly focuses on bimanual cooperative tasks of a single object. We build up 11 tasks in total, 8 for multi-task training and 3 for testing. The average success rate of stage one m_1 and trajectory tracking rate m_2 shown in Table 3 demonstrate the effectiveness and generalizability of BiDexHD in collaborative bimanual manipulation tasks. Refer to our Appendix for details.

Metrics (%)	BiDexHD-IPPO	BiDexHD-IPPO+DAgger
Train m_1	93.67	90.98
Train m_2	86.75	80.49
Test New m_1	80.31	88.62
Test New m_2	53.47	65.99

Table 3: The metrics of ARCTIC dataset.

Conclusion & Limitations

We introduce a novel approach to learning diverse bimanual dexterous manipulation skills that utilizes human demonstrations. Our BiDexHD automatically constructs bimanual manipulation tasks from existing datasets and employs a teacher-student learning approach for a vision-based policy that can tackle similar tasks. Experimentals demonstrate that BiDexHD facilitates robust RL training and policy distillation, achieves high performance across 6 categories of TACO unseen tool-using tasks through zero-shot generalization and transfers to ARCTIC collaborative tasks. In future work, exploring strategies to achieve more precise spatial and temporal tracking is a valuable future direction. Additionally, incorporating a wider variety of real-world tasks, such as deformable object manipulation and bimanual handover, could reveal further potential in dynamic collaborative manipulation scenarios with bimanual hands.

Acknowledgements

This work was supported by NSFC under Grant 62450001 and 62476008. The authors would like to thank the reviewers for their valuable comments and advice.

References

Arunachalam, S. P.; Silwal, S.; Evans, B.; and Pinto, L. 2023. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *2023 ieee international conference on robotics and automation (icra)*, 5954–5961. IEEE.

Bai, Y.; Yu, W.; and Liu, C. K. 2016. Dexterous manipulation of cloth. In *Computer Graphics Forum*.

Chen, T.; Tippur, M.; Wu, S.; Kumar, V.; Adelson, E.; and Agrawal, P. 2023. Visual dexterity: In-hand dexterous manipulation from depth. In *Icml workshop on new frontiers in learning, control, and dynamical systems*.

Chen, T.; Xu, J.; and Agrawal, P. 2022. A system for general in-hand object re-orientation. In *Conference on Robot Learning*, 297–307. PMLR.

Chen, Z.; Chen, S.; Schmid, C.; and Laptev, I. 2024. ViViDex: Learning Vision-based Dexterous Manipulation from Human Videos. *arXiv preprint arXiv:2404.15709*.

Cheng, X.; Li, J.; Yang, S.; Yang, G.; and Wang, X. 2024. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*.

Dasari, S.; Gupta, A.; and Kumar, V. 2023. Learning dexterous manipulation from exemplar object trajectories and pregrasps. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 3889–3896. IEEE.

De Witt, C. S.; Gupta, T.; Makoviichuk, D.; Makoviychuk, V.; Torr, P. H.; Sun, M.; and Whiteson, S. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*.

Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M. J.; and Hilliges, O. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12943–12954.

Fu, Z.; Zhao, Q.; Wu, Q.; Wetzstein, G.; and Finn, C. 2024. HumanPlus: Humanoid Shadowing and Imitation from Humans. *arXiv preprint arXiv:2406.10454*.

Gao, J.; Tao, Z.; Jaquier, N.; and Asfour, T. 2024. Bi-KVIL: Keypoints-based Visual Imitation Learning of Bimanual Manipulation Tasks. *arXiv preprint arXiv:2403.03270*.

Gbagbe, K. F.; Cabrera, M. A.; Alabbas, A.; Alyunes, O.; Lykov, A.; and Tsetserukou, D. 2024. Bi-VLA: Vision-Language-Action Model-Based System for Bimanual Robotic Dexterous Manipulations. *arXiv preprint arXiv:2405.06039*.

Grannen, J.; Wu, Y.; Vu, B.; and Sadigh, D. 2023. Stabilize to act: Learning to coordinate for bimanual manipulation. In *Conference on Robot Learning*, 563–576. PMLR.

Handa, A.; Van Wyk, K.; Yang, W.; Liang, J.; Chao, Y.-W.; Wan, Q.; Birchfield, S.; Ratliff, N.; and Fox, D. 2020. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 9164–9170. IEEE.

He, T.; Luo, Z.; He, X.; Xiao, W.; Zhang, C.; Zhang, W.; Kitani, K.; Liu, C.; and Shi, G. 2024. OmniH2O: Universal and Dexterous Human-to-Humanoid Whole-Body Teleoperation and Learning. *arXiv preprint arXiv:2406.08858*.

Hou, Y. C.; Sahari, K. S. M.; and How, D. N. T. 2019. A review on modeling of flexible deformable object for dexterous robotic manipulation. *International Journal of Advanced Robotic Systems*, 16(3): 1729881419848894.

Huang, B.; Chen, Y.; Wang, T.; Qin, Y.; Yang, Y.; Atanasov, N.; and Wang, X. 2023. Dynamic handover: Throw and catch with bimanual hands. *arXiv preprint arXiv:2309.05655*.

Huang, Z.; Yuan, H.; Fu, Y.; and Lu, Z. 2024. Efficient residual learning with mixture-of-experts for universal dexterous grasping. *arXiv preprint arXiv:2410.02475*.

Jia, X.; Blessing, D.; Jiang, X.; Reuss, M.; Donat, A.; Lioutikov, R.; and Neumann, G. 2024. Towards diverse behaviors: A benchmark for imitation learning with human demonstrations. *arXiv preprint arXiv:2402.14606*.

Li, S.; Huang, Z.; Chen, T.; Du, T.; Su, H.; Tenenbaum, J. B.; and Gan, C. 2023. Dexdeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics. *arXiv preprint arXiv:2304.03223*.

Lin, T.; Yin, Z.-H.; Qi, H.; Abbeel, P.; and Malik, J. 2024. Twisting lids off with two hands. *arXiv preprint arXiv:2403.02338*.

Liu, I.; Arthur, C.; He, S.; Seita, D.; and Sukhatme, G. 2024a. VoxAct-B: Voxel-Based Acting and Stabilizing Policy for Bimanual Manipulation. *arXiv preprint arXiv:2407.04152*.

Liu, Q.; Cui, Y.; Ye, Q.; Sun, Z.; Li, H.; Li, G.; Shao, L.; and Chen, J. 2023. DexRepNet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3153–3160. IEEE.

Liu, Y.; Yang, H.; Si, X.; Liu, L.; Li, Z.; Zhang, Y.; Liu, Y.; and Yi, L. 2024b. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21740–21751.

Luo, H.; Feng, Y.; Zhang, W.; Zheng, S.; Wang, Y.; Yuan, H.; Liu, J.; Xu, C.; Jin, Q.; and Lu, Z. 2025. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*.

Luo, Z.; Cao, J.; Christen, S.; Winkler, A.; Kitani, K.; and Xu, W. 2024. Grasping Diverse Objects with Simulated Humanoids. *arXiv preprint arXiv:2407.11385*.

Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A.; et al. 2021. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*.

Mandikal, P.; and Grauman, K. 2021. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, 6169–6176. IEEE.

Mandikal, P.; and Grauman, K. 2022. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, 651–661. PMLR.

Odesanmi, G. A.; Wang, Q.; and Mai, J. 2023. Skill learning framework for human–robot interaction and manipulation tasks. *Robotics and Computer-Integrated Manufacturing*, 79: 102444.

Qi, H.; Kumar, A.; Calandra, R.; Ma, Y.; and Malik, J. 2023. In-hand object rotation via rapid motor adaptation. In *Conference on Robot Learning*, 1722–1732. PMLR.

Qin, Y.; Wu, Y.-H.; Liu, S.; Jiang, H.; Yang, R.; Fu, Y.; and Wang, X. 2022. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, 570–587. Springer.

Qin, Y.; Yang, W.; Huang, B.; Van Wyk, K.; Su, H.; Wang, X.; Chao, Y.-W.; and Fox, D. 2023. AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System. In *Robotics: Science and Systems*.

Razali, H.; and Demiris, Y. 2023. Action-conditioned generation of bimanual object manipulation sequences. In *Proceedings of the AAAI conference on artificial intelligence*.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6).

Ross, S.; Gordon, G.; and Bagnell, D. 2011. 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*, 627–635. JMLR Workshop and Conference Proceedings.

Schmeckpeper, K.; Rybkin, O.; Daniilidis, K.; Levine, S.; and Finn, C. 2020. Reinforcement learning with videos: Combining offline observations with interaction. *arXiv preprint arXiv:2011.06507*.

Shao, L.; Migimatsu, T.; Zhang, Q.; Yang, K.; and Bohg, J. 2021. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14): 1419–1434.

Shaw, K.; Agarwal, A.; and Pathak, D. 2023. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. *arXiv preprint arXiv:2309.06440*.

Shaw, K.; Bahl, S.; and Pathak, D. 2023. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, 654–665. PMLR.

Sindhupathiraja, S. R.; Ullah, A. A.; Delamare, W.; and Hasan, K. 2024. Exploring Bi-Manual Teleportation in Virtual Reality. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 754–764. IEEE.

Sivakumar, A.; Shaw, K.; and Pathak, D. 2022. Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube. *arXiv preprint arXiv:2202.10448*.

Smith, L.; Dhawan, N.; Zhang, M.; Abbeel, P.; and Levine, S. 2019. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*.

Wan, W.; Geng, H.; Liu, Y.; Shan, Z.; Yang, Y.; Yi, L.; and Wang, H. 2023. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3891–3902.

Wang, C.; Shi, H.; Wang, W.; Zhang, R.; Fei-Fei, L.; and Liu, C. K. 2024a. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*.

Wang, S.; Liu, X.; Wang, C. C.; and Liu, J. 2024b. Physics-aware iterative learning and prediction of saliency map for bimanual grasp planning. *Computer Aided Geometric Design*, 111: 102298.

Xu, Y.; Wan, W.; Zhang, J.; Liu, H.; Shan, Z.; Shen, H.; Wang, R.; Geng, H.; Weng, Y.; Chen, J.; et al. 2023. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4737–4746.

Yin, Z.-H.; Huang, B.; Qin, Y.; Chen, Q.; and Wang, X. 2023. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*.

Yu, D.; Xu, H.; Chen, Y.; Ren, Y.; and Pan, J. 2024. BiKC: Keypose-Conditioned Consistency Policy for Bimanual Robotic Manipulation. *arXiv preprint arXiv:2406.10093*.

Yuan, H.; Bai, Y.; Fu, Y.; Zhou, B.; Feng, Y.; Xu, X.; Zhan, Y.; Karlsson, B. F.; and Lu, Z. 2025. Being-0: A Humanoid Robotic Agent with Vision-Language Models and Modular Skills. *arXiv preprint arXiv:2503.12533*.

Yuan, H.; Zhou, B.; Fu, Y.; and Lu, Z. 2024. Cross-embodiment dexterous grasping with reinforcement learning. *arXiv preprint arXiv:2410.02479*.

Zhan, X.; Yang, L.; Zhao, Y.; Mao, K.; Xu, H.; Lin, Z.; Li, K.; and Lu, C. 2024. 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*, 445–456.

Zhang, H.; Christen, S.; Fan, Z.; Zheng, L.; Hwangbo, J.; Song, J.; and Hilliges, O. 2024a. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, 235–246. IEEE.

Zhang, T.; Li, D.; Li, Y.; Zeng, Z.; Zhao, L.; Sun, L.; Chen, Y.; Wei, X.; Zhan, Y.; Li, L.; et al. 2024b. Empowering Embodied Manipulation: A Bimanual-Mobile Robot Manipulation Dataset for Household Tasks. *arXiv preprint arXiv:2405.18860*.

Zhou, B.; Li, K.; Jiang, J.; and Lu, Z. 2023. Learning from visual observation via offline pretrained state-to-go transformer. *Advances in Neural Information Processing Systems*, 36: 59585–59605.

Zhou, B.; Zhan, Y.; Zhang, Z.; and Lu, Z. 2025. MEgoHand: Multimodal Egocentric Hand-Object Interaction Motion Generation. *arXiv preprint arXiv:2505.16602*.

Zhou, B.; Zhang, Z.; Wang, J.; and Lu, Z. 2024. NOLO: Navigate Only Look Once. *arXiv preprint arXiv:2408.01384*.