2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3315-4364-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVPR52734.2025.00656

MANIPTRANS: Efficient Dexterous Bimanual Manipulation Transfer via Residual Learning

Kailin Li[1] Puhao Li[1,2] Tengyu Liu[1] Yuyang Li[1,3] Siyuan Huang[1]
[1]State Key Laboratory of General Artificial Intelligence, BIGAI
[2]Department of Automation, Tsinghua University [3]Institute for Artificial Intelligence, Peking University
https://maniptrans.github.io

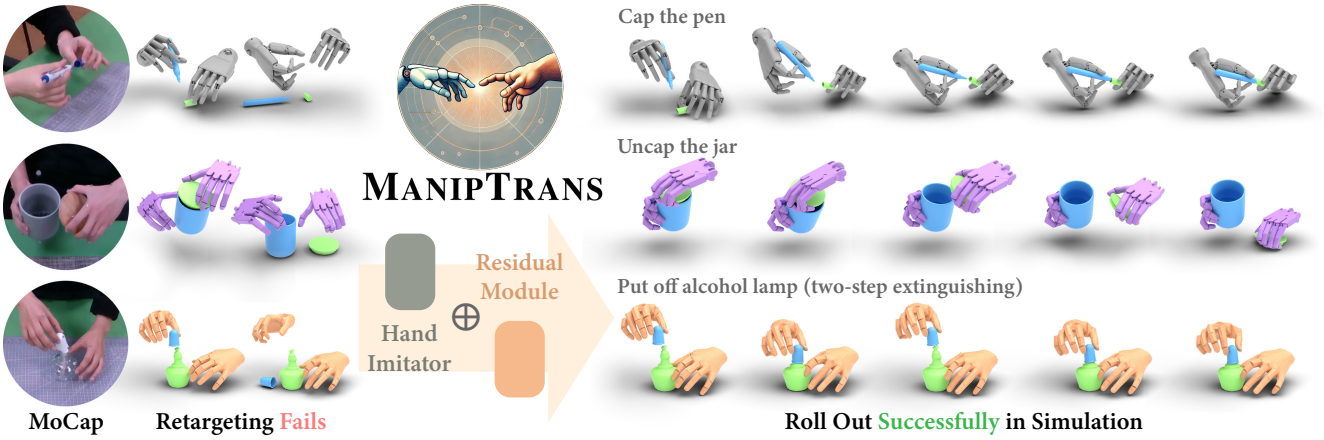


Figure 1. **MANIPTRANS for Bimanual Dexterous Manipulations.** Retargeting methods often struggle with transferring MoCap data to physically plausible motions, while our **MANIPTRANS** efficiently produces task-compliant, physically accurate motions. It also generalizes across embodiments like Inspire hands [3], Shadow hands [1], and articulated MANO hands [27, 94].

Abstract

Human hands play a central role in interacting, motivating increasing research in dexterous robotic manipulation. Data-driven embodied AI algorithms demand precise, large-scale, human-like manipulation sequences, which are challenging to obtain with conventional reinforcement learning or real-world teleoperation. To address this, we introduce **MANIPTRANS***, a novel two-stage method for efficiently transferring human bimanual skills to dexterous robotic hands in simulation.* **MANIPTRANS** *first pre-trains a generalist trajectory imitator to mimic hand motion, then fine-tunes a specific residual module under interaction constraints, enabling efficient learning and accurate execution of complex bimanual tasks. Experiments show that* **MANIPTRANS** *surpasses state-of-the-art methods in success rate, fidelity, and efficiency. Leveraging* **MANIPTRANS***, we transfer multiple hand-object datasets to robotic hands, creating* **DEXMANIPNET***, a large-scale dataset featuring previously unexplored tasks like pen capping and bottle unscrewing.* **DEXMANIPNET** *comprises 3.3K episodes of robotic manipulation and is easily extensible, facilitating further policy training for dexterous hands and enabling real-world deployments.*

1. Introduction

Embodied AI (EAI) has advanced rapidly in recent years, with increasing efforts to enable AI-driven embodiments to interact with physical or virtual environments. Just as human hands are pivotal for interaction, much research in EAI focuses on dexterous robotic hand manipulation [4, 16–22, 41, 46, 50, 56, 57, 61, 63, 64, 66, 68, 70, 73, 75, 79, 80, 102, 111, 113, 116, 127, 128]. Achieving human-like proficiency in complex bimanual tasks holds significant research value and is crucial for progress toward general AI.

Thus, the rapid acquisition of precise, large-scale, and human-like dexterous manipulation sequences for data-driven embodied agents training [11, 12, 25, 81, 130] becomes increasingly urgent. Some studies use reinforcement learning (RL) [52, 97] to explore and generate dexterous hand actions [27, 67, 75, 109, 119, 132, 133], while others collect human-robot paired data through teleoperation [26, 44, 45, 80, 101, 111, 125]. Both methods are limited: traditional RL requires carefully designed, task-specific reward functions [76, 132], restricting scalability

2575-7075/25/$31.00 ©2025 IEEE
DOI 10.1109/CVPR52734.2025.00656

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

and task complexity, while teleoperation is labor-intensive and costly, yielding only embodiment-specific datasets.

A promising solution is to transfer human manipulation actions to dexterous robotic hands in simulated environments via imitation learning [69, 78, 91, 110, 136]. This approach offers several advantages. First, imitating human manipulation trajectories creates naturalistic hand-object interactions, enabling more fluid and human-like motions. Second, abundant motion-capture (MoCap) datasets [10, 14, 32, 37, 39, 55, 60, 71, 72, 105, 117, 123, 131] and hand pose estimation techniques [13, 43, 65, 85, 106, 118, 120–122, 124] makes extracting operational knowledge from human demonstrations easily accessible [91, 100]. Third, simulations provide a cost-effective validation, offering a shortcut to real-world robot deployment [41, 44, 49].

Yet, achieving precise and efficient transfer is non-trivial. As shown in Fig. 1, morphological differences between human and robotic hands lead to direct pose retargeting suboptimal. Additionally, although MoCap data is relatively accurate, error accumulation can still lead to critical failures during high-precision tasks. Moreover, bimanual manipulation introduces a high-dimensional action space, significantly increasing the difficulty of efficient policy learning. Consequently, most pioneering work generally stops at single-hand grasping and lifting tasks [27, 109, 119, 132], leaving complex bimanual activities—such as unscrewing a bottle or capping a pen—largely unexplored.

In this paper, we propose a simple but efficient method, **ManipTrans**, which facilitates the transfer of hand manipulation skills—especially bimanual actions—to dexterous robotic hands in simulation, enabling accurate tracking of reference motions. *Our key insight is to treat the transfer as a two-stage process: a pre-training trajectory imitation stage focusing on hand motion alone, followed by a specific action fine-tuning stage that meets interaction constraints.* Specifically, we design a robust generalist model that learns to accurately mimic human finger motions with resilience to noise. Based on this initial imitation, we then introduce a residual learning module [47, 49, 51, 104] that incrementally refines the robot's actions, focusing on two key aspects: 1) ensuring stable contact with object surfaces under physical constraints, enabling effective object manipulation, and 2) coordinating both hands to ensure precise, high-fidelity execution of complex bimanual operations.

The advantages of this design are threefold: 1) In the first stage, focusing on dynamic hand mimicry with large-scale pretraining **effectively mitigates morphological differences**. 2) Building on this advantage, the second stage concentrates on tracking bimanual object interactions, **enabling precise capture of subtle movements and facilitating natural, high-fidelity manipulation**. 3) It **significantly reduces action space complexity** by decoupling human hand motion imitation from physics-based object interaction constraints, thus improving training efficiency.

Building on this framework, **ManipTrans** corrects arbitrary, noisy hand MoCap data into physically plausible motion without predefined stages (*e.g.*, "approaching-grasping-manipulation") or task-specific reward engineering. We, therefore, validate its effectiveness and efficiency across a range of complex single- and bimanual manipulations, including articulated object handling [32, 34, 60, 105, 131]. Using **ManipTrans**, we transfer several representative hand-object manipulation datasets [60, 131] to dexterous robotic hands in the Isaac Gym simulation [77], constructing the **DexManipNet** dataset, which achieves marked improvements in motion fidelity and compliance. Currently, **DexManipNet** comprises 3.3K episodes and 1.34 million frames of robotic hand manipulation, covering previously unexplored tasks such as pen capping, bottle cap unscrewing, and chemical experimentation.

We experimentally demonstrate that **ManipTrans** outperforms baseline methods in both motion precision and transfer success rate. Notably, it surpasses prior state-of-the-art (SOTA) approaches in transfer efficiency, even on a personal computer. To evaluate its extensibility, we conducted cross-embodiment experiments applying **ManipTrans** to dexterous hands with varying degrees of freedom (DoFs) and morphologies, achieving consistent performance with minimal additional effort. Furthermore, we replay **DexManipNet**'s bimanual trajectories on real-world devices, demonstrating agile and natural dexterous manipulation that, to the best of our knowledge, has not been achieved by previous RL- or teleoperation-based methods. Finally, we benchmark **DexManipNet** using several imitation learning frameworks, underscoring its value to the research community.

In summary, our contributions are as follows:

- We introduce **ManipTrans**, a simple yet effective two-stage transfer framework that enables precise transfer of human bimanual manipulation to dexterous robotic hands in simulation, ensuring accurate tracking of both hand and object reference motions.
- Using this framework, we construct **DexManipNet**, a large-scale, high-quality dataset featuring a wide array of novel bimanual manipulation tasks with high precision and compliance. **DexManipNet** is extensible and serves as a valuable resource for future policy training.
- Our experiments show that **ManipTrans** outperforms previous SOTA methods. We further demonstrate its generalizability across various dexterous hand configurations and its feasibility for real-world deployment.

2. Related Works

Dexterous Manipulation via Human Demonstration
Learning manipulation skills from human demonstrations offers an intuitive and effective approach to transferring

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

human abilities to robots [6, 31, 126, 129]. Imitation learning has shown considerable promise in achieving this transfer [7, 23, 62, 69, 78, 87, 88, 107, 110, 136, 139]. Recent studies focus on learning RL policies guided by object trajectories [21, 22, 70, 75, 139]. QuasiSim [70] advances this approach by directly transferring reference hand motions to robotic hands via parameterized quasi-physical simulators. However, these methods are limited to simpler tasks and are computationally intensive. More recently, tailored solutions using task-specific reward functions have been developed for challenging tasks like bimanual lip-twisting [66, 68]. In contrast, our method enables efficient learning of complex manipulation tasks without task-specific reward engineering.

Dexterous Hand Datasets Object manipulation is fundamental for embodied agents. Numerous MANO-based [94] hand-object interaction datasets exist [9, 10, 14, 28, 32, 36, 37, 39, 40, 42, 53, 55, 58–60, 71, 72, 91, 98, 105, 117, 123, 131, 138, 140]. However, these datasets often prioritize pose alignment with 2D images while neglecting physical constraints, limiting their applicability for robotic training. Teleoperation methods [26, 44, 45, 90, 111, 115, 125, 137] collect human-to-robot hand matching data online using AR/VR systems [15, 24, 30, 50, 84] or vision-based MoCap [92, 111, 112] for real-time data acquisition and correction with humans in the loop. However, teleoperation is labor-intensive and time-consuming, and the absence of tactile feedback often yields stiff, unnatural actions, hindering fine-grained manipulation. In contrast, our method enables offline transfer of human demonstrations to robots. Our **DexManipNet** offers a large, easily expandable collection of human demonstration episodes.

Residual Learning Due to the sample inefficiency and time-consuming nature of RL training, residual policy learning [51, 96, 104], which incrementally refines action control, is widely adopted to enhance efficiency and stability. In dexterous hand manipulation, various studies explore residual strategies tailored to specific tasks [5, 21, 29, 38, 96, 116, 135, 136]. For instance, [38] integrates user input during residual policy training, while [49] learns corrective actions from human demonstrations. GraspGF [116] employs a pre-trained score-based generative model as a base, and [21] decomposes the imitation task into wrist following and finger motion control, integrating a residual wrist control policy. Additionally, [47] constructs a mixture-of-experts system [48] using residual learning, and DexH2R [136] applies residual learning directly to retargeted robotic hand actions. Our method differs from these approaches by pre-training a finger motion imitation model that incorporates additional dynamic information, followed by fine-tuning a residual policy to adapt to task-specific physical constraints. This approach is more efficient and generalizable across various manipulation tasks.

3. Method

We provide an overview of our method in Fig. 2. Given reference human hand–object interaction trajectories, our goal is to learn a policy that enables dexterous robotic hands to accurately replicate these trajectories in simulation while satisfying the task's semantic manipulation constraints. To this end, we propose a two-stage framework: the first stage trains a general hand trajectory imitation model, and the second stage employs a residual model to refine the initial coarse motion into task-compliant actions.

3.1. Preliminaries

Without loss of generality, we formulate the manipulation transfer problem in a complex bimanual setting, where the left and right dexterous hands, $\boldsymbol{d} = \{d_l, d_r\}$, aim to replicate the behavior of human hands, $\boldsymbol{h} = \{h_l, h_r\}$, which interact with two objects, $\boldsymbol{o} = \{o_l, o_r\}$, in a cooperative manner (e.g., in a pen-capping task where one hand holds the cap while the other grips the pen body). The reference trajectories from human demonstrations are defined as $\boldsymbol{\mathcal{T}_h} = \{\boldsymbol{\tau_h^t}\}_{t=1}^T$ and $\boldsymbol{\mathcal{T}_o} = \{\boldsymbol{\tau_o^t}\}_{t=1}^T$, where T represents the total number of frames. The trajectory $\boldsymbol{\tau_h}$ for each hand includes the wrist's 6-DoF pose $\boldsymbol{w_h} \in \mathbb{SE}(3)$, the linear and angular velocities $\dot{\boldsymbol{w}}_{\boldsymbol{h}} = \{\boldsymbol{v_h}, \boldsymbol{u_h}\}$, and the finger joint positions $\boldsymbol{j_h} \in \mathbb{R}^{F\times 3}$ defined by MANO [94], along with their respective velocities $\dot{\boldsymbol{j}}_{\boldsymbol{h}} = \{\boldsymbol{v_j}, \boldsymbol{u_j}\}$; here, F denotes the number of hand keypoints, including the fingertips. Similarly, the object trajectory $\boldsymbol{\tau_o}$ for each object includes its 6-DoF pose $\boldsymbol{p_o} \in \mathbb{SE}(3)$ and the corresponding linear and angular velocities $\dot{\boldsymbol{p}}_{\boldsymbol{o}} = \{\boldsymbol{v_o}, \boldsymbol{u_o}\}$. To reduce spatial complexity, we normalize all translations relative to the dexterous hand's wrist position while preserving the original rotations to maintain the correct gravity direction.

We model this problem as an implicit Markov Decision Process (MDP) $\mathcal{M} = \langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{A}}, \mathbf{T}, \mathbf{R}, \gamma \rangle$, where $\boldsymbol{\mathcal{S}}$ represents the state space, $\boldsymbol{\mathcal{A}}$ the action space, $\mathbf{T}$ the transition dynamics, $\mathbf{R}$ the reward function, and γ the discount factor. The action for each dexterous hand at time t, denoted as $\boldsymbol{a}^t \in \boldsymbol{\mathcal{A}}$, comprises the target positions of each dexterous hand's joint $\boldsymbol{a_q^t} \in \mathbb{R}^K$ for proportional-derivative (PD) control, and the 6-DoF force $\boldsymbol{a_w^t} \in \mathbb{R}^6$ applied to the robotic wrist, similar to prior work [47, 109, 119], where K denotes the total number of robotic hand revolute joints (*i.e.* the DoF).

Our approach divides the transfer process into two stages: 1) a pre-trained hand-only trajectory imitation model $\boldsymbol{\mathcal{I}}$, and 2) a residual module $\boldsymbol{\mathcal{R}}$ that fine-tunes the coarse actions to ensure task compliance. The state at time t is defined separately for each stage as $\boldsymbol{s_\mathcal{I}^t} \in \boldsymbol{\mathcal{S}_\mathcal{I}}$ and $\boldsymbol{s_\mathcal{R}^t} \in \boldsymbol{\mathcal{S}_\mathcal{R}}$, with corresponding reward functions $r_\mathcal{I}^t = \mathbf{R}(\boldsymbol{s_\mathcal{I}^t}, \boldsymbol{a_\mathcal{I}^t})$ and $r_\mathcal{R}^t = \mathbf{R}(\boldsymbol{s_\mathcal{R}^t}, \boldsymbol{a_\mathcal{R}^t})$ as described in Sec. 3.2 and Sec. 3.3. For both stages, we employ proximal policy optimization (PPO) [97] to maximize the discounted reward $\mathbb{E}\left[\sum_{t=1}^T \gamma^{t-1} r_{\text{stage}}^t\right]$, following previous methods [19, 87].

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

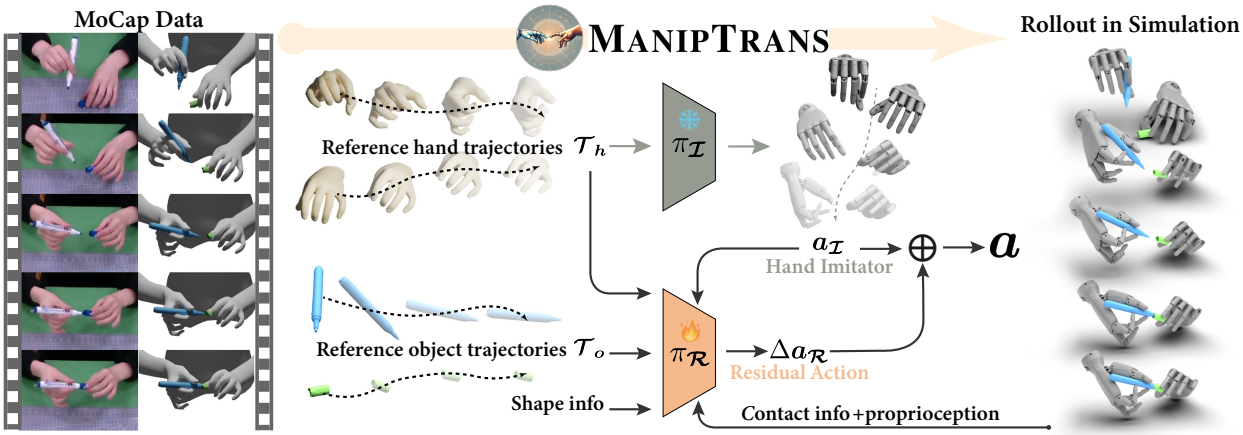


Figure 2. **Our ManipTrans Pipeline.** We first pre-train a hand motion imitation model with large-scale human demonstrations, then fine-tune a residual policy to adapt to task-specific physical constraints.

3.2. Hand Trajectory Imitating

In this stage, our objective is to learn a general hand trajectory imitation model, $\mathcal{I}$, capable of accurately replicating detailed human finger motions. The state for each dexterous hand at time t is defined as $\boldsymbol{s}_{\mathcal{I}}^t = \{\boldsymbol{\tau}_{\boldsymbol{h}}^t, \boldsymbol{s}_{\text{prop}}^t\}$, which includes the target hand trajectory $\boldsymbol{\tau}_{\boldsymbol{h}}^t$ and the current proprioception $\boldsymbol{s}_{\text{prop}}^t = \{\boldsymbol{q}_{\boldsymbol{d}}^t, \dot{\boldsymbol{q}}_{\boldsymbol{d}}^t, \boldsymbol{w}_{\boldsymbol{d}}^t, \dot{\boldsymbol{w}}_{\boldsymbol{d}}^t\}$. Here, $\boldsymbol{q}_{\boldsymbol{d}}^t$ and $\boldsymbol{w}_{\boldsymbol{d}}^t$ denote the joint angles and wrist poses, respectively, along with their corresponding velocities. We aim to train the policy $\pi_{\mathcal{I}}(\boldsymbol{a}^t|\boldsymbol{s}_{\mathcal{I}}^t, \boldsymbol{a}^{t-1})$ using RL to determine the actions $\boldsymbol{a}_{\mathcal{I}}^t$.

Reward Functions. The reward function $r_{\mathcal{I}}^t$ is designed to encourage the dexterous hands to track the reference hand trajectory $\boldsymbol{\tau}_{\boldsymbol{h}}^t$ while ensuring stability and smoothness. It comprises three components: 1) *Wrist tracking reward* r_{wrist}^t: This reward minimizes the difference: $\boldsymbol{w}_{\boldsymbol{d}}^t \ominus \boldsymbol{w}_{\boldsymbol{h}}^t$ and $\dot{\boldsymbol{w}}_{\boldsymbol{d}}^t - \dot{\boldsymbol{w}}_{\boldsymbol{h}}^t$, $\ominus$ denotes the difference in $\mathbb{SE}(3)$ space. 2) *Finger imitation reward* r_{finger}^t: This component encourages the dexterous hand to closely follow the reference finger joint positions. We manually select F finger keypoints on the dexterous hand corresponding to the MANO model, denoted as $\boldsymbol{j}_{\boldsymbol{d}}$. The weights w_f and decay rates λ_f are empirically set to emphasize the fingertips, particularly those of the thumb, index, and middle fingers. The parameters are in the Appx. This design helps mitigate the impact of morphological differences between human and robotic hands:

$$r_{\text{finger}}^t = \sum_{f=1}^{F} w_f \cdot \exp\left(-\lambda_f \|\boldsymbol{j}_{\boldsymbol{d}_f}^t - \boldsymbol{j}_{\boldsymbol{h}_f}^t\|_2^2\right) \quad (1)$$

3) *Smoothness Reward* r_{smooth}^t: To alleviate jerky motions, we introduce a smoothness reward that penalizes the power exerted on each joint, defined as the element-wise product of joint velocities and torques, similar to the approach in [74]. The total reward is defined as: $r_{\mathcal{I}}^t = w_{\text{wrist}} \cdot r_{\text{wrist}}^t + w_{\text{finger}} \cdot r_{\text{finger}}^t + w_{\text{smooth}} \cdot r_{\text{smooth}}^t$.

Training Strategy. Decoupling hand imitation from object interaction offers additional benefits; specifically, $\pi_{\mathcal{I}}$ does not require challenging-to-acquire manipulation data. We train the policy using hand-only datasets, including existing hand motion collections [14, 36, 60, 105, 131, 134, 141] and synthetic data generated via interpolation [103]. To balance training data between the left and right hands, we mirror these datasets; training time and additional details are provided in the Appx. For efficiency, we employ reference state initialization (RSI) and early termination [86, 87]. If the dexterous hand keypoints $\boldsymbol{j}_{\boldsymbol{d}}$ deviate beyond a threshold ϵ_{finger}, the episode terminates early and resets to a randomly sampled MoCap state. We also utilize curriculum learning [8], gradually reducing ϵ_{finger} to encourage broad exploration initially, then focusing on fine-grained finger control.

3.3. Residual Learning for Interaction

Building on the pre-trained $\pi_{\mathcal{I}}$, we use a residual module $\mathcal{R}$ to refine coarse actions and satisfy task-specific constraints.

State Space Expansion for Interaction. To account for interactions between the dexterous hands and objects, we expand the state space beyond the hand-related states $\boldsymbol{s}_{\mathcal{I}}^t$ by incorporating additional interaction-related information. First, we compute the convex hull [114] of the object meshes $\boldsymbol{o}$ from MoCap data to generate the collidable object $\hat{\boldsymbol{o}}$ in the simulation environment. To manipulate the object along the reference $\boldsymbol{\mathcal{T}}_{\boldsymbol{o}}$, we include the object's position $\boldsymbol{p}_{\hat{\boldsymbol{o}}}$ (relative to the wrist position $\boldsymbol{w}_{\boldsymbol{d}}$) and velocities $\dot{\boldsymbol{p}}_{\hat{\boldsymbol{o}}}$, center of mass $\boldsymbol{m}_{\hat{\boldsymbol{o}}}$, and gravitational force vector $\boldsymbol{G}_{\hat{\boldsymbol{o}}}$. To better encode the object's shape, we utilize the BPS representation [89]. Additionally, for enhancing perception, we encode the spatial relationship between the hands and the object using the distance metric: $\boldsymbol{D}(\boldsymbol{j}_{\boldsymbol{d}}^t, \boldsymbol{p}_{\hat{\boldsymbol{o}}}^t) = \|\boldsymbol{j}_{\boldsymbol{d}}^t - \boldsymbol{p}_{\hat{\boldsymbol{o}}}^t\|_2^2$, measuring the squared Euclidean distance between the dexterous hand keypoints and the object's position. Furthermore, we explicitly include the contact force $\boldsymbol{C}$ obtained from the simulation, capturing the interaction between the fingertips and the object's surface. This tactile feedback is critical for stable grasping and manipulation, ensuring precise control dur-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

ing complex tasks. In summary, the expanded interaction state for the residual module is defined as: $\boldsymbol{s}^t_{\text{interact}} = \{\boldsymbol{\tau}^t_{\boldsymbol{o}}, \boldsymbol{p}^t_{\hat{\boldsymbol{o}}}, \dot{\boldsymbol{p}}^t_{\hat{\boldsymbol{o}}}, \boldsymbol{m}^t_{\hat{\boldsymbol{o}}}, \boldsymbol{G}^t_{\hat{\boldsymbol{o}}}, \text{BPS}(\hat{\boldsymbol{o}}), \boldsymbol{D}(\boldsymbol{j}^t_{\boldsymbol{d}}, \boldsymbol{p}^t_{\hat{\boldsymbol{o}}}), \boldsymbol{C}^t\}$.

Residual Actions Combining Strategy. Given the combined state $\boldsymbol{s}^t_{\boldsymbol{\mathcal{R}}} = \boldsymbol{s}^t_{\boldsymbol{\mathcal{I}}} \cup \boldsymbol{s}^t_{\text{interact}}$, our goal is to learn residual actions $\Delta\boldsymbol{a}^t_{\boldsymbol{\mathcal{R}}}$ that refine the initial imitation actions $\boldsymbol{a}^t_{\boldsymbol{\mathcal{I}}}$ to ensure task compliance. During each step of the manipulation episode, we first sample the imitation action $\boldsymbol{a}^t_{\boldsymbol{\mathcal{I}}} \sim \pi_{\boldsymbol{\mathcal{I}}}(\boldsymbol{a}^t|\boldsymbol{s}^t_{\boldsymbol{\mathcal{I}}}, \boldsymbol{a}^{t-1})$. Conditioned on this action, we then sample the residual correction $\Delta\boldsymbol{a}^t_{\boldsymbol{\mathcal{R}}} \sim \pi_{\boldsymbol{\mathcal{R}}}(\Delta\boldsymbol{a}^t|\boldsymbol{s}^t_{\boldsymbol{\mathcal{R}}}, \boldsymbol{a}^t_{\boldsymbol{\mathcal{I}}}, \boldsymbol{a}^{t-1})$. The final action is computed as: $\boldsymbol{a}^t = \boldsymbol{a}^t_{\boldsymbol{\mathcal{I}}} + \Delta\boldsymbol{a}^t_{\boldsymbol{\mathcal{R}}}$, where the residual action is added element-wise. The resulting action $\boldsymbol{a}^t$ is then clipped to adhere to the dexterous hand's joint limits. At the start of training, since the dexterous hand movements already approximate the reference hand trajectory, the residual actions are expected to be close to zero. This initialization helps prevent model collapse and accelerates convergence. We achieve this by initializing the residual module with a zero-mean Gaussian distribution and employing a warm-up strategy to gradually activate its training.

Reward Functions. Our objective is to efficiently transfer human bimanual manipulation skills to dexterous robotic hands in a task-agnostic manner. To this end, we avoid task-specific reward engineering, which, although beneficial for individual tasks, can limit generalization. Therefore, our reward design remains simple and general. In addition to the hand imitation reward $r^t_{\boldsymbol{\mathcal{I}}}$ discussed in Sec. 3.2, we introduce two additional components: 1) *Object following reward* r^t_{object}: Minimizes positional and velocity differences between the simulated object and its reference trajectory, specifically $\boldsymbol{p}^t_{\hat{\boldsymbol{o}}} \ominus \boldsymbol{p}^t_{\boldsymbol{o}}$ and $\dot{\boldsymbol{p}}^t_{\hat{\boldsymbol{o}}} - \dot{\boldsymbol{p}}^t_{\boldsymbol{o}}$. 2) *Contact force reward* r^t_{contact}: Encourages appropriate contact force when the hand-object distance in the MoCap dataset is below a specified threshold ξ_{c}. The reward is defined as:

$$r^t_{\text{contact}} = w_{\text{c}} \cdot \exp\left(\frac{-\lambda_{\text{c}}}{\sum_{f=1}^{F} \boldsymbol{C}^t_{\boldsymbol{d}_f} \cdot \mathbb{1}\left(\boldsymbol{D}(\boldsymbol{j}^t_{\boldsymbol{h}_f}, \boldsymbol{p}^t_{\boldsymbol{o}} \cdot \boldsymbol{o}) < \xi_{\text{c}}\right)}\right) \tag{2}$$

where $\boldsymbol{D}(\boldsymbol{j}^t_{\boldsymbol{h}_f}, \boldsymbol{p}^t_{\boldsymbol{o}} \cdot \boldsymbol{o})$ represents the minimum distance between the fingertip $\boldsymbol{h}_f$ and the transformed object surface, $\mathbb{1}(\cdot)$ is the indicator function, and $\boldsymbol{C}^t_{\boldsymbol{d}_f}$ denotes the contact force at the fingertip. The weight w_{c} and decay rate λ_{c} are empirically set to balance the reward function. The total reward for the residual stage is defined as $r^t_{\boldsymbol{\mathcal{R}}} = r^t_{\boldsymbol{\mathcal{I}}} + w_{\text{object}} \cdot r^t_{\text{object}} + w_{\text{contact}} \cdot r^t_{\text{contact}}$.

Training Strategy. Inspired by prior work [70, 82, 83] that utilizes quasi-physical simulators to relax constraints during training and avoid local minima, we introduce a relaxation mechanism in the residual learning stage. Unlike [70], which employs custom simulations, we adjust the physical constraints directly within the Isaac Gym environment [77] to enhance training efficiency. Specifically, we initially set the gravitational constant $\mathcal{G}$ to zero and the friction coefficient $\mathcal{F}$ to a high value. This setup allows the robotic hands to, early in training, grip objects firmly and efficiently align with reference trajectories. As training progresses, we gradually restore $\mathcal{G}$ to its true value and reduce $\mathcal{F}$ to a suitable value to approximate real interactions. Similar to the imitation stage, we adopt RSI, early termination, and curriculum learning strategies. Each episode initializes the robotic hands by randomly selecting a non-colliding near-object state from the preprocessed trajectory. During training, if the object's pose $\boldsymbol{p}^t_{\hat{\boldsymbol{o}}}$ deviates beyond a predefined threshold ϵ_{object}, the episode is terminated early. We progressively reduce ϵ_{object} to encourage more precise object manipulation. Additionally, we introduce a contact termination condition: if MoCap data indicates a firm grasp by the human hands (i.e., $\boldsymbol{D}(\boldsymbol{j}^t_{\boldsymbol{h}_f}, \boldsymbol{p}^t_{\boldsymbol{o}} \cdot \boldsymbol{o}) < \xi_{\text{t}}$, where ξ_{t} is the termination threshold), the contact force $\boldsymbol{C}^t_{\boldsymbol{d}_f}$ must be non-zero. Failure to meet this condition results in early termination. This mechanism ensures the agent learns to control contact forces, promoting stable object manipulation.

3.4. DexManipNet Dataset

Using **ManipTrans**, we generate **DexManipNet**, derived from two representative large-scale hand-object interaction datasets: FAVOR [60] and OakInk-V2 [131]. FAVOR employs VR-based teleoperation with human-in-the-loop corrections, focusing on foundational tasks like object rearrangement. In contrast, OakInk-V2 utilizes optical tracking-based motion capture, targeting more complex interactions such as pen capping and bottle unscrewing.

Due to the lack of standardization in dexterous robotic hands, we adopt the Inspire Hand [3] as our primary platform for its high dexterity, stability, cost-effectiveness, and extensive prior use [24, 35, 50]. To address the complexity of bimanual tasks, we employ a simulated 12-DoF configuration of the Inspire Hand, enhancing flexibility compared to its real-world 6-DoF mechanism. We demonstrate **ManipTrans**'s adaptability to other robotic hands and real-world deployment in Sec. 4.4 and Sec. 4.5.

Our **DexManipNet** encompasses 61 diverse and challenging tasks as defined in [131], comprising 3.3K episodes of robotic hand manipulation over 1.2K objects, totaling 1.34 million frames, including $\sim$ 600 sequences involving complex bimanual tasks. Each episode executes precisely in the Isaac Gym simulation [77]. In comparison, a recent dataset generated via automated augmentation [50] includes only 60 source human demonstrations across 9 tasks.

4. Experiments

In experiments, we describe the dataset setup and metrics (Sec. 4.1), followed by implementation details (Sec. 4.2). We then compare **ManipTrans** with SOTA methods

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

(Sec. 4.3), demonstrate cross-embodiment generalization (Sec. 4.4), validate real-world deployment (Sec. 4.5), conduct ablation studies (Sec. 4.6), and benchmark **DexManipNet** for learning manipulation policies (Sec. 4.7).

4.1. Datasets and Metrics

Datasets For quantitative evaluation, we use the official validation dataset of OakInk-V2 [131], approximately half of which consists of bimanual tasks. To assess transfer capabilities, we manually select MoCap sequences that meet task completeness and semantic relevance, filtering them to durations of 4–20 seconds and downsampling to 60 fps. We exclude sequences involving deformable or oversized objects, resulting in $\sim$ 80 episodes. For qualitative evaluation, we also incorporate the GRAB [105], FAOVR [60], and ARCTIC [32] datasets to demonstrate our advantages.

Metrics To evaluate **ManipTrans** in terms of manipulation precision, task compliance, and transfer efficiency, we introduce the following metrics. These are adapted from [70] but are more stringent due to the complexity of our bimanual tasks: 1) Per-frame Average Object Rotation and Translation Error: $E_r = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{p}_{\text{rot}}{}_{\hat{\boldsymbol{o}}}^{t} \cdot (\boldsymbol{p}_{\text{rot}}{}_{\boldsymbol{o}}^{t})^{-1})$ and $E_t = \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{p}_{\text{tsl}}{}_{\hat{\boldsymbol{o}}}^{t} - \boldsymbol{p}_{\text{tsl}}{}_{\boldsymbol{o}}^{t}\|_2^2$. Here, $\boldsymbol{p}_{\text{rot}}$ and $\boldsymbol{p}_{\text{tsl}}$ are the rotation and translation components of the 6-DoF pose $\boldsymbol{p}$, respectively. Errors E_r and E_t are reported in degrees and centimeters. 2) Mean Per-Joint Position Error (in cm): $E_j = \frac{1}{T \cdot F}\sum_{t=1}^{T}\sum_{f=1}^{F}\|\boldsymbol{j}_{\boldsymbol{d}_f}^{t} - \boldsymbol{j}_{\boldsymbol{h}_f}^{t}\|_2^2$. This metric measures the average error in the positions of the hand joints. 3) Mean Per-Fingertip Position Error (in cm): $E_{ft} = \frac{1}{T \cdot M}\sum_{t=1}^{T}\sum_{ft=1}^{M}\|\boldsymbol{t}_{\boldsymbol{d}_{ft}}^{t} - \boldsymbol{t}_{\boldsymbol{h}_{ft}}^{t}\|_2^2$. This metric evaluates the mimicry quality of fingertip $\boldsymbol{t}$ motions, accounting for morphological differences between human and robotic hands. Here, M equals 5 for single-hand tasks and 10 for bimanual tasks. 4) Success Rate (SR): A tracking attempt is deemed successful if E_r, E_t, E_j, and E_{ft} are all below the specified thresholds: $30°$, 3 cm, 8 cm, and 6 cm, respectively. For bimanual tasks, the trajectory is considered failed if either hand fails to meet these conditions, making the success criterion stricter compared to single-hand tasks.

4.2. Implementation Details

In **ManipTrans**, we manually selected $F = 21$ keypoints on each dexterous robotic hand, corresponding to the fingertips, palm, and phalangeal positions on the human hand, to mitigate the morphological differences. Details on keypoint selection and weight coefficients w for reward terms are provided in Appx. For training, we use a curriculum learning strategy. The initial threshold ϵ_{finger} is set to 6 cm and decays to 4 cm. Object alignment thresholds ϵ_{object} start at $90°$ and 6 cm for rotation and translation, gradually decreasing to $30°$ and 2 cm. We train both the imitation module $\boldsymbol{\mathcal{I}}$ and residual module $\boldsymbol{\mathcal{R}}$ using the Actor-Critic PPO algorithm [97], with a training horizon of 32 frames, a mini-batch size of 1024, and a discount factor $\gamma = 0.99$. Optimization employs Adam [54] with an initial learning rate of 5×10^{-4} and a decay scheduler. All experiments are run in Isaac Gym [77], simulating 4096 environments at a time step of $1/60$ s on a personal computer equipped with an NVIDIA RTX 4090 GPU and an Intel i9-13900KF CPU.

4.3. Evaluations

As discussed in Sec. 2, dexterous hand manipulation advances rapidly, with previous approaches differing in problem formulations and task definitions. To offer a comprehensive and fair comparison, we evaluate two categories of methods—RL-combined and optimization-based—to demonstrate **ManipTrans**'s accuracy and efficiency.

Comparison with RL-Combined Methods Due to the lack of publicly available code for prior RL-combined methods, we reimplement representative approaches: 1) *RL-Only* exploration using only trajectory-following rewards, employing the PPO algorithm to train the robotic hand from scratch based on [27]; 2) *Retarget + Residual* learning, applying residual action to retargeted robotic hand poses obtained via alignment between human and robot keypoints [92]. As a naive baseline, we also include the *Retarget-Only* method—retargeting without any learning.

As shown in Tab. 1, our method outperforms all baselines across multiple metrics, demonstrating superior precision in both single- and bimanual tasks. These results confirm that our two-stage transfer framework effectively captures subtle finger motions and object interactions, leading to high task success rates and motion fidelity.

We find that the *Retarget-Only* baseline is nearly infeasible due to the complexity of the dexterous hand action space and error accumulation. The *RL-Only* baseline performs suboptimally since exploration from scratch is time-consuming and reduces motion precision. Compared to the *Retarget + Residual* baseline, our method—leveraging a pre-trained hand imitation model—demonstrates improved control capabilities, enabling more accurate manipulation aligned with the reference trajectory. Notably, the Retargeting method often causes collisions in contact-rich scenarios, resulting in instability during residual policy training. We further study **ManipTrans**'s robustness and time cost in Appx. Fig. 3 shows the qualitative results

Methods	$E_r \downarrow$	$E_t \downarrow$	$E_j \downarrow$	$E_{ft} \downarrow$	$SR \uparrow$
Retarget-Only	N/A	N/A	N/A	N/A	4.6 / 0.0
RL-Only	9.72	1.23	2.96	2.38	34.3 / 12.1
Retarget + Residual	11.58	0.79	2.54	1.74	47.8 / 13.9
ManipTrans	**8.60**	**0.49**	**2.15**	**1.36**	**58.1 / 39.5**

Table 1. **Quantitative Comparisons with RL-Combined Baselines.** The first four metrics are computed only on successfully rolled-out sequences. The SR includes the separated transfer success rates for single/bimanual tasks. The error scores on *Retarget-Only* are not available since it hardly works.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

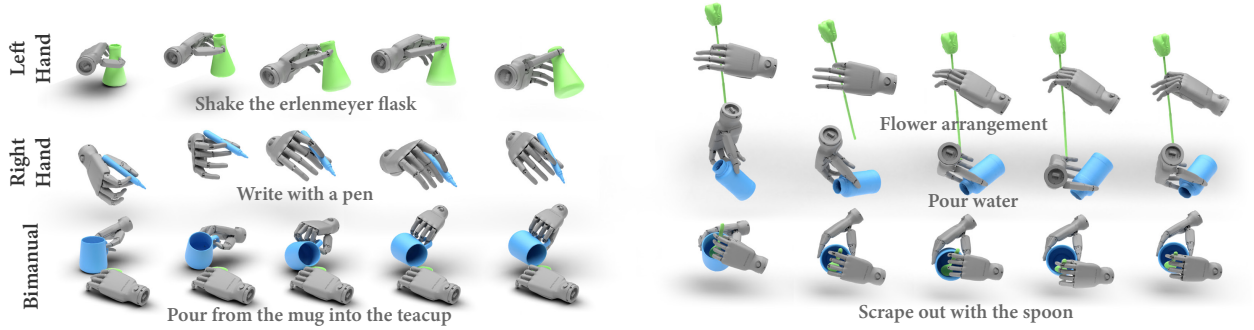


Figure 3. **Qualitative Results of ManipTrans.** We showcase the transfer results using the Inspire left and right hands on both single-hand tasks (top two rows) and bimanual tasks (bottom row) from the OakInk-V2 [131] dataset. Notably, the dexterous hands successfully manipulate delicate and slim objects, such as a pen and a flower stem.

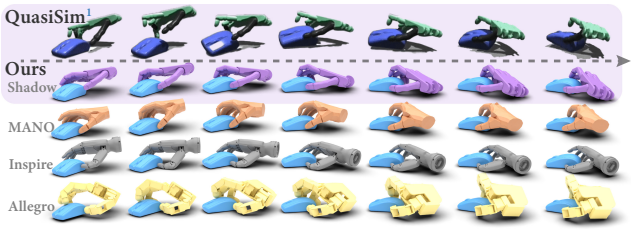


Figure 4. **Qualitative Comparison with QuasiSim [70]. ManipTrans** produces more natural motion of the Shadow hand (purple region) and is applicable to other dexterous hands.

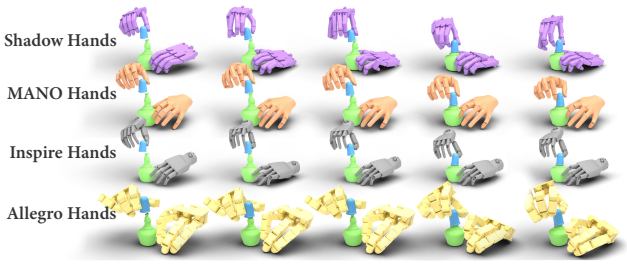


Figure 5. **Cross Embodiments Results**: Putting off Alcohol lamp.

on seldom-explored tasks, highlighting the natural and precision of **ManipTrans** transferring human manipulation skills. Additional details and more qualitative results applying our method to articulated objects are provided in Appx.

Comparison with Optimization-Based Method QuasiSim [70] optimizes over customized simulations to track human motions. Currently, their full pipeline has not yet been released, and their "randomly" selected validation set is not available. Thus, a direct quantitative comparison is not feasible. Therefore, we provide a qualitative comparison in Fig. 4, demonstrating **ManipTrans**'s ability to transfer human motions to the Shadow Hand in a setting similar to QuasiSim's, but with more stable contacts and smoother motions. Notably, due to our two-stage design, for an unseen single-hand manipulation trajectory of 60 frames ("rotating a mouse"), our method requires $\sim$ 15 minutes of training to achieve robust results, compared to QuasiSim's $\sim$ 40 hours of optimization[1], highlighting **ManipTrans**'s significant efficiency.

4.4. Cross-Embodiments Validation

We demonstrate **ManipTrans**'s extensibility across various dexterous hand embodiments. As described in Sec. 3, the imitation module $\mathcal{I}$ addresses hand keypoint tracking, while the residual module $\mathcal{R}$ captures physical interactions between fingertips and objects. Our framework is embodiment-agnostic since it relies solely on the correspondence between human fingers and robotic joints, allowing adaptation to different dexterous hands with minimal effort. We evaluate **ManipTrans** on the Shadow Hand [1], articulated MANO hand [27, 94], Inspire Hand [3], and Allegro Hand [2], which have varying DoFs: K = 22, 22, 12, and 16, respectively. Without altering network hyperparameters or reward weights, **ManipTrans** achieves consistent, fluid, and precise performance across all embodiments in both single-hand tasks (Fig. 4) and bimanual tasks (Fig. 5). Additional details on the Allegro Hand—a robotic hand with only four fingers—are provided in Appx.

4.5. Real-World Deployment

As illustrated in Fig. 6, we conduct experiments using two 7-DoF Realman arms [93] and a pair of upgraded Inspire Hands (same configuration yet adding tactile sensors). To bridge the gap between the simulated 12-DoF robotic hands and the 6-DoF real hardware, we employ a fitting-based method that optimizes the joint angles $\boldsymbol{q}_{\tilde{\boldsymbol{d}}} \in \mathbb{R}^6$ of the real robots (denoted as $\tilde{\cdot}$) for fingertip alignment, formulated as: $\operatorname{argmin}_{\boldsymbol{q}_{\tilde{\boldsymbol{d}}}} \frac{1}{T \cdot M} \sum_{t=1}^{T} \sum_{ft=1}^{M} \|\boldsymbol{t}^t_{\boldsymbol{d}_{ft}} - \boldsymbol{t}^t_{\tilde{\boldsymbol{d}}_{ft}}\|_2^2$ with an additional temporal smoothness loss: $L_{\text{smooth}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\boldsymbol{q}^{t+1}_{\tilde{\boldsymbol{d}}} - \boldsymbol{q}^t_{\tilde{\boldsymbol{d}}}\|_2^2$. We control the arms by solving inverse kinematics to align the arms' flanges with the dexterous hands' wrists $\boldsymbol{w}_{\boldsymbol{d}}$. During replay, we do not enforce strict temporal alignment, as the real robots cannot always operate as quickly as human hands.

Fig. 6 showcases dexterous manipulation that, to the best

[1]Results shown in QuasiSim's Appx and its official repository: https://github.com/Meowuu7/QuasiSim

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

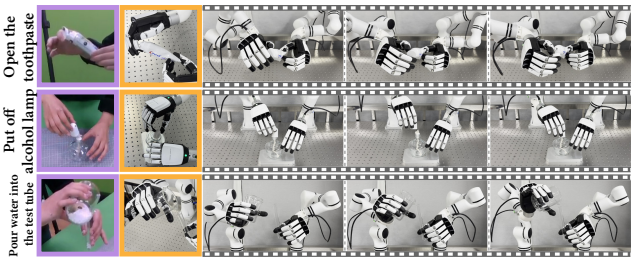
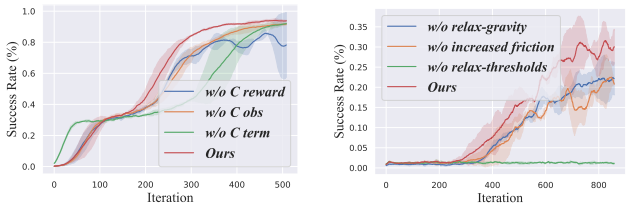


Figure 6. **Real-world bimanual manipulation deployment.** Purple box: human hand motion; orange box: close-up of dexterous hands. More results are on the website. (Zoom in for details.)



(a) Tactile ablations training curve. (b) Curve on training strategies.

Figure 7. **Training Curve of Ablation Studies.** We assess tactile feedback in contact-rich tasks (*e.g.*, turning off a lamp) and curriculum learning in complex ones (*e.g.*, capping a pen).

Methods	IBC [33]	BET [99]	DP-UNet [25]	DP-Trans [25]
SR	4.69%	9.69%	**18.44%**	14.69%

Table 2. **Imitating Learning on Bottle Rearrangement Task.**

of our knowledge, has not previously been achieved. For example, in "opening the toothpaste", the left hand stably holds the tube while the right hand's thumb and index finger flexibly pop open the tiny cap—motions challenging to capture via teleoperation. This underscores the potential of our method for future real-world policy learning.

4.6. Abalation Studies

Tactile Information as Auxiliary Input In Sec. 3.3, we integrate tactile information, specifically the contact force $\boldsymbol{C}$, into the pipeline in three ways: (1) as an observation input, (2) as a reward component to encourage contact, and (3) as a condition for early termination. Ablation studies (Fig. 7a) labeled *w/o* $\boldsymbol{C}$ *obs*, *w/o* $\boldsymbol{C}$ *reward*, and *w/o* $\boldsymbol{C}$ *term* demonstrate that including $\boldsymbol{C}$ in the reward function improves task success rates, and treating $\boldsymbol{C}$ as an observation accelerates convergence. We also find that omitting $\boldsymbol{C}$ as a termination condition seems to enhance initial training performance but lowers overall convergence speed, highlighting the importance of stable contact in task completion.

Training Strategy We begin training with a curriculum learning strategy that includes (1) relaxing gravity effects, (2) increasing friction influence, and (3) relaxing thresholds ϵ_{finger} and ϵ_{object}. Ablation studies (Fig. 7b), labeled *w/o relax-gravity*, *w/o increased friction*, and *w/o relax-thresholds*, show that for precise, complex bimanual motions, ignoring gravity and using high friction coefficients in the early stages accelerate convergence and achieve higher overall SR. Without initial relaxation of the threshold constraints, the network may fail to converge entirely.

4.7. DexManipNet for Policy Learning

To benchmark **DexManipNet**'s potential, we evaluate representative imitation learning methods on a fundamental policy learning task: rearrangement. Specifically, we focus on *moving a bottle to a goal position*. Given the bottle's current and goal 6D poses, the environment state (including obstacles on the table), and the dexterous hand's proprioception, the policy generates a sequence of robotic hand actions to pick up the bottle and place it at the target.

We evaluate four representative imitation learning methods: two regression-based behavior cloning approaches—IBC [33] and BET [99]—and two diffusion policy methods [25] with UNet [95] and Transformer [108] backbones. Each policy is trained on 85% of the 140 sequences involving the bottle rearrangement task in **DexManipNet** and evaluated on the remaining 15%. We perform 20 rollouts per sequence. A rollout is considered successful if the object's final position is within 10 cm of the goal. Further details are provided in Appx.

As shown in Tab. 2, all methods perform suboptimally due to the task's difficulty and the complexity of the dexterous hand action space. Regression-based behavior cloning approaches, in particular, suffer from error accumulation. These results highlight the inherent challenges of dexterous manipulation tasks, which require precise finger control and effective object manipulation. We hope that **DexManipNet** will facilitate advancements in this domain.

5. Conclusion and Discussion

ManipTrans is a two-stage framework that efficiently transfers human manipulation skills to dexterous robotic hands. By decoupling hand motion imitation from object interaction via residual learning, **ManipTrans** overcomes morphological differences and complex task challenges, ensuring high-fidelity motions and efficient training. Experiments demonstrate that **ManipTrans** surpasses SOTA methods in motion precision and computational efficiency, while also exhibiting cross-embodiment adaptability and feasibility for real-world deployment. Furthermore, the extensible **DexManipNet** establishes a new benchmark to advance progress in embodied AI.

Discussion and Limitations Although **ManipTrans** successfully handles most MoCap data, some sequences cannot be transferred effectively. We attribute this to two main reasons: 1) excessive noise in interaction poses and 2) insufficiently accurate object models for simulation, particularly for articulated objects. Enhancing **ManipTrans**'s robustness and generating physically plausible object models are valuable directions for future research.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

References

[1] ShadowRobot. https://www.shadowrobot.com/dexterous-hand-series, 2005. 1, 7

[2] Allegro Hands. https://www.allegrohand.com, 2013. 7

[3] Inspire Hands. https://en.inspire-robots.com/product-category/the-dexterous-hands, 2019. 1, 5, 7

[4] Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. In *CoRL*, 2023. 1

[5] Minttu Alakuijala, Gabriel Dulac-Arnold, Julien Mairal, Jean Ponce, and Cordelia Schmid. Residual reinforcement learning from demonstrations. *arXiv preprint arXiv:2106.08050*, 2021. 3

[6] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 2009. 3

[7] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *ICRA*, 2023. 3

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 2009. 4

[9] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 3

[10] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020. 2, 3

[11] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1

[12] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1

[13] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 2

[14] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 2, 3, 4

[15] Sirui Chen, Chen Wang, Kaden Nguyen, Li Fei-Fei, and C Karen Liu. Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback. *arXiv preprint arXiv:2410.08464*, 2024. 3

[16] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. In *CoRL*, 2022. 1

[17] Tao Chen, Megha Tippur, Siyang Wu, Vikash Kumar, Edward Adelson, and Pulkit Agrawal. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 2023.

[18] Tao Chen, Eric Cousineau, Naveen Kuppuswamy, and Pulkit Agrawal. Vegetable peeling: A case study in constrained dexterous manipulation. *arXiv preprint arXiv:2407.07884*, 2024.

[19] Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *NeurIPS*, 2022. 3

[20] Yuanpei Chen, Chen Wang, Li Fei-Fei, and C Karen Liu. Sequential dexterity: Chaining dexterous policies for long-horizon manipulation. *arXiv preprint arXiv:2309.00987*, 2023.

[21] Yuanpei Chen, Chen Wang, Yaodong Yang, and Karen Liu. Object-centric dexterous manipulation from human motion data. In *CoRL*, 2024. 3

[22] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. Vividex: Learning vision-based dexterous manipulation from human videos. *arXiv preprint arXiv:2404.15709*, 2024. 1, 3

[23] Zoey Qiuyu Chen, Karl Van Wyk, Yu-Wei Chao, Wei Yang, Arsalan Mousavian, Abhishek Gupta, and Dieter Fox. Dextransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations. *arXiv preprint arXiv:2209.14284*, 2022. 3

[24] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024. 3, 5

[25] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *IJRR*, 2023. 1, 8

[26] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024. 1, 3

[27] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *CVPR*, 2022. 1, 2, 6, 7

[28] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 3

[29] Todor Davchev, Kevin Sebastian Luck, Michael Burke, Franziska Meier, Stefan Schaal, and Subramanian Ramamoorthy. Residual learning from demonstration: Adapting dmps for contact-rich manipulation. *RA-L*, 2022. 3

[30] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. *arXiv preprint arXiv:2407.03162*, 2024. 3

[31] Peter Englert and Marc Toussaint. Learning manipulation skills from a single demonstration. *IJRR*, 2018. 3

[32] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 2, 3, 6

[33] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *CoRL*, 2022. 8

[34] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. *arXiv preprint arXiv:2412.04244*, 2024. 2

[35] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024. 5

[36] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. Dart: Articulated hand model with diverse accessories and rich textures. *NeurIPS*, 2022. 3, 4

[37] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 2, 3

[38] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In *IROS*, 2020. 3

[39] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2, 3

[40] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, 2022. 3

[41] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makovichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, et al. Dextreme: Transfer of agile in-hand manipulation from simulation to reality. In *ICRA*. IEEE, 2023. 1, 2

[42] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 3

[43] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 2

[44] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024. 1, 2, 3

[45] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024. 1, 3

[46] Binghao Huang, Yuanpei Chen, Tianyu Wang, Yuzhe Qin, Yaodong Yang, Nikolay Atanasov, and Xiaolong Wang. Dynamic handover: Throw and catch with bimanual hands. *CoRL*, 2023. 1

[47] Ziye Huang, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Efficient residual learning with mixture-of-experts for universal dexterous grasping. *arXiv preprint arXiv:2410.02475*, 2024. 2, 3

[48] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 1991. 3

[49] Yunfan Jiang, Chen Wang, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Transic: Sim-to-real policy transfer by learning from online correction. In *CoRL*, 2024. 2, 3

[50] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024. 1, 3, 5

[51] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *ICRA*, 2019. 2, 3

[52] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 1996. 1

[53] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. *arXiv preprint arXiv:2401.10232*, 2024. 3

[54] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[55] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 2, 3

[56] Haoming Li, Qi Ye, Yuchi Huo, Qingtao Liu, Shijian Jiang, Tao Zhou, Xiang Li, Yang Zhou, and Jiming Chen. Tpgp: Temporal-parametric optimization with deep grasp prior for dexterous motion planning. In *ICRA*, 2024. 1

[57] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. *arXiv preprint arXiv:2410.11792*, 2024. 1

[58] Kailin Li, Lixin Yang, Haoyu Zhen, Zenan Lin, Xinyu Zhan, Licheng Zhong, Jian Xu, Kejian Wu, and Cewu Lu. Chord: Category-level hand-held object reconstruction via shape deformation. In *ICCV*, 2023. 3

[59] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. In *ECCV*, 2024.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

[60] Kailin Li, Lixin Yang, Zenan Lin, Jian Xu, Xinyu Zhan, Yifei Zhao, Pengxiang Zhu, Wenxiong Kang, Kejian Wu, and Cewu Lu. Favor: Full-body ar-driven virtual object rearrangement guided by instruction text. *AAAI*, 2024. 2, 3, 4, 5, 6

[61] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *ICRA*, 2023. 1

[62] Sizhe Li, Zhiao Huang, Tao Chen, Tao Du, Hao Su, Joshua B Tenenbaum, and Chuang Gan. Dexdeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics. *ICLR*, 2023. 3

[63] Yuyang Li, Bo Liu, Yiran Geng, Puhao Li, Yaodong Yang, Yixin Zhu, Tengyu Liu, and Siyuan Huang. Grasp multiple objects with one hand. *RA-L*, 2024. 1

[64] Davide Liconti, Yasunori Toshimitsu, and Robert Katzschmann. Leveraging pretrained latent representations for few-shot imitation learning on a dexterous robotic hand. *arXiv preprint arXiv:2404.16483*, 2024. 1

[65] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2

[66] Toru Lin, Zhao-Heng Yin, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Twisting lids off with two hands. *arXiv preprint arXiv:2403.02338*, 2024. 1, 3

[67] Qingtao Liu, Yu Cui, Qi Ye, Zhengnan Sun, Haoming Li, Gaofeng Li, Lin Shao, and Jiming Chen. Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. In *IROS*, 2023. 1

[68] Qingtao Liu, Qi Ye, Zhengnan Sun, Yu Cui, Gaofeng Li, and Jiming Chen. Masked visual-tactile pre-training for robot manipulation. In *ICRA*, 2024. 1, 3

[69] Wenhai Liu, Junbo Wang, Yiming Wang, Weiming Wang, and Cewu Lu. Force-centric imitation learning with force-motion capture system for contact-rich manipulation. *arXiv preprint arXiv:2410.07554*, 2024. 2, 3

[70] Xueyi Liu, Kangbo Lyu, Jieqiong Zhang, Tao Du, and Li Yi. Parameterized quasi-physical simulators for dexterous manipulations transfer. In *ECCV*, 2024. 1, 3, 5, 6, 7

[71] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022. 2, 3

[72] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. *arXiv preprint arXiv:2401.08399*, 2024. 2, 3

[73] Haoran Lu, Ruihai Wu, Yitong Li, Sijie Li, Ziyu Zhu, Chuanruo Ning, Yan Shen, Longzan Luo, Yuanpei Chen, and Hao Dong. Garmentlab: A unified simulation and benchmark for garment manipulation. In *NeurIPS*, 2024. 1

[74] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023. 4

[75] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids. *arXiv preprint arXiv:2407.11385*, 2024. 1, 3

[76] Zhengyi Luo, Jiashun Wang, Kangni Liu, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, et al. Smplolympics: Sports environments for physically simulated humanoids. *arXiv preprint arXiv:2407.00187*, 2024. 1

[77] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 2, 5, 6

[78] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, 2018. 2, 3

[79] Xiaofeng Mao, Gabriele Giudici, Claudio Coppola, Kaspar Althoefer, Ildar Farkhatdinov, Zhibin Li, and Lorenzo Jamone. Dexskills: Skill segmentation using haptic data for learning autonomous long-horizon robotic manipulation tasks. *arXiv preprint arXiv:2405.03476*, 2024. 1

[80] Ji-Heon Oh, Ismael Espinoza, Danbi Jung, and Tae-Seong Kim. Bimanual long-horizon manipulation via temporal-context transformer rl. *RA-L*, 2024. 1

[81] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 1

[82] Tao Pang and Russ Tedrake. A convex quasistatic time-stepping scheme for rigid multibody systems with contact and friction. In *ICRA*, 2021. 5

[83] Tao Pang, HJ Terry Suh, Lujie Yang, and Russ Tedrake. Global planning for contact-rich manipulation via local smoothing of quasi-dynamic contact models. *IEEE Transactions on Robotics*, 2023. 5

[84] Younghyo Park, Jagdeep Singh Bhatia, Lars Ankile, and Pulkit Agrawal. Dexhub and dart: Towards internet scale robot data collection. *arXiv preprint arXiv:2411.02214*, 2024. 3

[85] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024. 2

[86] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM TOG*, 2018. 4

[87] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM TOG*, 2021. 3, 4

[88] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM TOG*, 2022. 3

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

[89] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, 2019. 4

[90] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *RA-L*, 2022. 3

[91] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*, 2022. 2, 3

[92] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *RSS*, 2023. 3, 6

[93] Realman Robotics. RM Series. https://www.realman-robotics.com/rm-series123, 2010. 7

[94] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017. 1, 3, 7

[95] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 8

[96] Gerrit Schoettler, Ashvin Nair, Jianlan Luo, Shikhar Bahl, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Deep reinforcement learning for industrial insertion tasks with visual inputs and natural rewards. In *IROS*, 2020. 3

[97] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1, 3, 6

[98] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, 2022. 3

[99] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *NeurIPS*, 2022. 8

[100] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *CoRL*, 2023. 2

[101] Kenneth Shaw, Yulong Li, Jiahui Yang, Mohan Kumar Srirama, Ray Liu, Haoyu Xiong, Russell Mendonca, and Deepak Pathak. Bimanual dexterity for complex tasks. *arXiv preprint arXiv:2411.13677*, 2024. 1

[102] Qijin She, Shishun Zhang, Yunfan Ye, Min Liu, Ruizhen Hu, and Kai Xu. Learning cross-hand policies for high-dof reaching and grasping. *ECCV*, 2024. 1

[103] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, 1985. 4

[104] Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018. 2, 3

[105] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 2, 3, 4, 6

[106] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 2

[107] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *SIGGRAPH ASIA*, 2024. 3

[108] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 8

[109] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *ICCV*, 2023. 1, 2, 3

[110] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 2, 3

[111] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024. 1, 3

[112] Jun Wang, Yuzhe Qin, Kaiming Kuang, Yigit Korkmaz, Akhilan Gurumoorthy, Hao Su, and Xiaolong Wang. Cyberdemo: Augmenting simulated human demonstration for real-world dexterous manipulation. In *CVPR*, 2024. 3

[113] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *ICRA*, 2023. 1

[114] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM TOG*, 2022. 4

[115] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. *arXiv preprint arXiv:2309.13037*, 2023. 3

[116] Tianhao Wu, Mingdong Wu, Jiyao Zhang, Yunchong Gan, and Hao Dong. Learning score-based grasping primitive for human-assisting dexterous grasping. *NeurIPS*, 2024. 1, 3

[117] Wei Xie, Zhipeng Yu, Zimeng Zhao, Binghui Zuo, and Yangang Wang. Hmdo: Markerless multi-view hand manipulation capture with deformable objects. *Graphical Models*, 2023. 2, 3

[118] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022. 2

[119] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *CVPR*, 2023. 1, 2, 3

[120] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE TPAMI*, 2023. 2

[121] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.

[122] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, 2022. 2

[123] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, 2022. 2, 3

[124] Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, and Cewu Lu. Poem: reconstructing hand in a point embedded multi-view stereo. In *CVPR*, 2023. 2

[125] Shiqi Yang, Minghuan Liu, Yuzhe Qin, Runyu Ding, Jialong Li, Xuxin Cheng, Ruihan Yang, Sha Yi, and Xiaolong Wang. Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. *arXiv preprint arXiv:2408.11805*, 2024. 1, 3

[126] Jianglong Ye, Jiashun Wang, Binghao Huang, Yuzhe Qin, and Xiaolong Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. *RA-L*, 2023. 3

[127] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *RSS*, 2023. 1

[128] Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning. *arXiv preprint arXiv:2410.02479*, 2024. 1

[129] Kevin Zakka, Philipp Wu, Laura Smith, Nimrod Gileadi, Taylor Howell, Xue Bin Peng, Sumeet Singh, Yuval Tassa, Pete Florence, Andy Zeng, et al. Robopianist: Dexterous piano playing with deep reinforcement learning. *CoRL*, 2023. 3

[130] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *RSS*, 2024. 1

[131] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *CVPR*, 2024. 2, 3, 4, 5, 6, 7

[132] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. *ECCV*, 2024. 1, 2

[133] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *3DV*, 2024. 1

[134] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *ICIP*, 2017. 4

[135] Xiang Zhang, Changhao Wang, Lingfeng Sun, Zheng Wu, Xinghao Zhu, and Masayoshi Tomizuka. Efficient sim-to-real transfer of contact-rich manipulation skills with online admittance residual learning. In *CoRL*, 2023. 3

[136] Shuqi Zhao, Xinghao Zhu, Yuxin Chen, Chenran Li, Xiang Zhang, Mingyu Ding, and Masayoshi Tomizuka. Dexh2r: Task-oriented dexterous manipulation from human to robots. *arXiv preprint arXiv:2411.04428*, 2024. 2, 3

[137] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 3

[138] Licheng Zhong, Lixin Yang, Kailin Li, Haoyu Zhen, Mei Han, and Cewu Lu. Color-neus: Reconstructing neural implicit surfaces with color. In *3DV*, 2024. 3

[139] Bohan Zhou, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations. *arXiv preprint arXiv:2410.02477*, 2024. 3

[140] Zehao Zhu, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. Contactart: Learning 3d interaction priors for category-level articulated object and hand poses estimation. *arXiv preprint arXiv:2305.01618*, 2023. 3

[141] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. 4

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:25 UTC from IEEE Xplore. Restrictions apply.