# H-RDT: Human Manipulation Enhanced Bimanual Robotic Manipulation

**Hongzhe Bi**[1,2], **Lingxuan Wu**[1], **Tianwei Lin**[2], **Hengkai Tan**[1],
**Zhizhong Su**[2], **Hang Su**[1], **Jun Zhu**[1]
[1]Dept. of Comp. Sci. and Tech., Institute for AI, BNRist Center, THBI Lab,
Tsinghua-Bosch Joint ML Center, Tsinghua University
[2]Horizon Robotics
bhz24@mails.tsinghua.edu.cn

## Abstract

Imitation learning for robotic manipulation faces a fundamental challenge: the scarcity of large-scale, high-quality robot demonstration data. Recent robotic foundation models often pre-train on cross-embodiment robot datasets to increase data scale, while they face significant limitations as the diverse morphologies and action spaces across different robot embodiments make unified training challenging. In this paper, we present **H-RDT** (**H**uman to **R**obotics **D**iffusion **T**ransformer), a novel approach that leverages human manipulation data to enhance robot manipulation capabilities. Our key insight is that large-scale egocentric human manipulation videos with paired 3D hand pose annotations provide rich behavioral priors that capture natural manipulation strategies and can benefit robotic policy learning. We introduce a two-stage training paradigm: (1) pre-training on large-scale egocentric human manipulation data, and (2) cross-embodiment fine-tuning on robot-specific data with modular action encoders and decoders. Built on a diffusion transformer architecture with 2B parameters, H-RDT uses flow matching to model complex action distributions. The modular design of action encoder and decoder components enables effective knowledge transfer from the unified human embodiment to diverse robot platforms through efficient fine-tuning. Extensive evaluations encompassing both simulation and real-world experiments, single-task and multitask scenarios, as well as few-shot learning and robustness assessments, demonstrate that H-RDT outperforms training from scratch and existing state-of-the-art methods, including $\pi_0$ and RDT, achieving significant improvements of **13.9%** and **40.5%** over training from scratch in simulation and real-world experiments, respectively. The results validate our core hypothesis that human manipulation data can serve as a powerful foundation for learning bimanual robotic manipulation policies.

## Introduction

Recent advances in robotic learning have been driven by specialized action policies like ACT (Zhao et al. 2023), Diffusion Policy (Chi et al. 2023), and 3D Diffusion Policy (Ze et al. 2024), as well as Vision-Language-Action (VLA) models such as RT-2 (Brohan et al. 2023), OpenVLA (Kim et al. 2024), RDT (Liu et al. 2024), $\pi_0$ (Black et al. 2024), and $\pi_{0.5}$ (Intelligence et al. 2025). However, these approaches face fundamental data collection challenges. Robot demonstration data relies heavily on teleoperation (Zhao et al. 2023; Aldaco et al. 2024), which requires expensive equipment and skilled operators, while advanced data collection systems like Universal Manipulation Interface (Chi et al. 2024) and motion capture setups (Wang et al. 2024; Xu et al. 2025) suffer from complex infrastructure requirements and inconsistent data quality that limit scalability.

Current VLA models typically employ cross-embodiment pre-training on robot datasets like Open X-Embodiment (O'Neill et al. 2024) and AgiBot World Colosseo (Bu et al. 2025). This approach faces two critical limitations: the diverse morphologies and action spaces across robot embodiments make unified training challenging, and existing robot datasets remain limited in scale with heterogeneous data quality across different collection setups. These constraints fundamentally limit the data availability and generalization capabilities needed for general-purpose robotic manipulation (Team et al. 2024; O'Neill et al. 2024).

In stark contrast, human manipulation behaviors represent a vast, readily accessible repository of demonstration data. The recent emergence of large-scale egocentric video datasets with detailed hand pose annotations, exemplified by EgoDex (Hoque et al. 2025) with its 829 hours of manipulation videos, offers unprecedented opportunities for learning rich behavioral priors. Human demonstrations naturally capture object affordances, manipulation strategies, and task decomposition patterns that could potentially serve as powerful inductive biases for robotic learning. Recent works have begun exploring this direction: EgoMimic (Kareer et al. 2024) employs co-training on human and robot data with egocentric video, while Humanoid Policy (HAT) (Qiu et al. 2025) uses differentiable retargeting for human-humanoid behavior modeling.

This paper introduces H-RDT (Human to Robotics Diffusion Transformer), a novel approach that systematically leverages large-scale egocentric human manipulation data to enhance robot manipulation capabilities. Our approach focuses on three specific aspects: **Data Scarcity:** We harness the abundance of human manipulation videos with 3D hand pose annotations to provide rich behavioral priors that capture natural manipulation strategies, object affordances, and

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

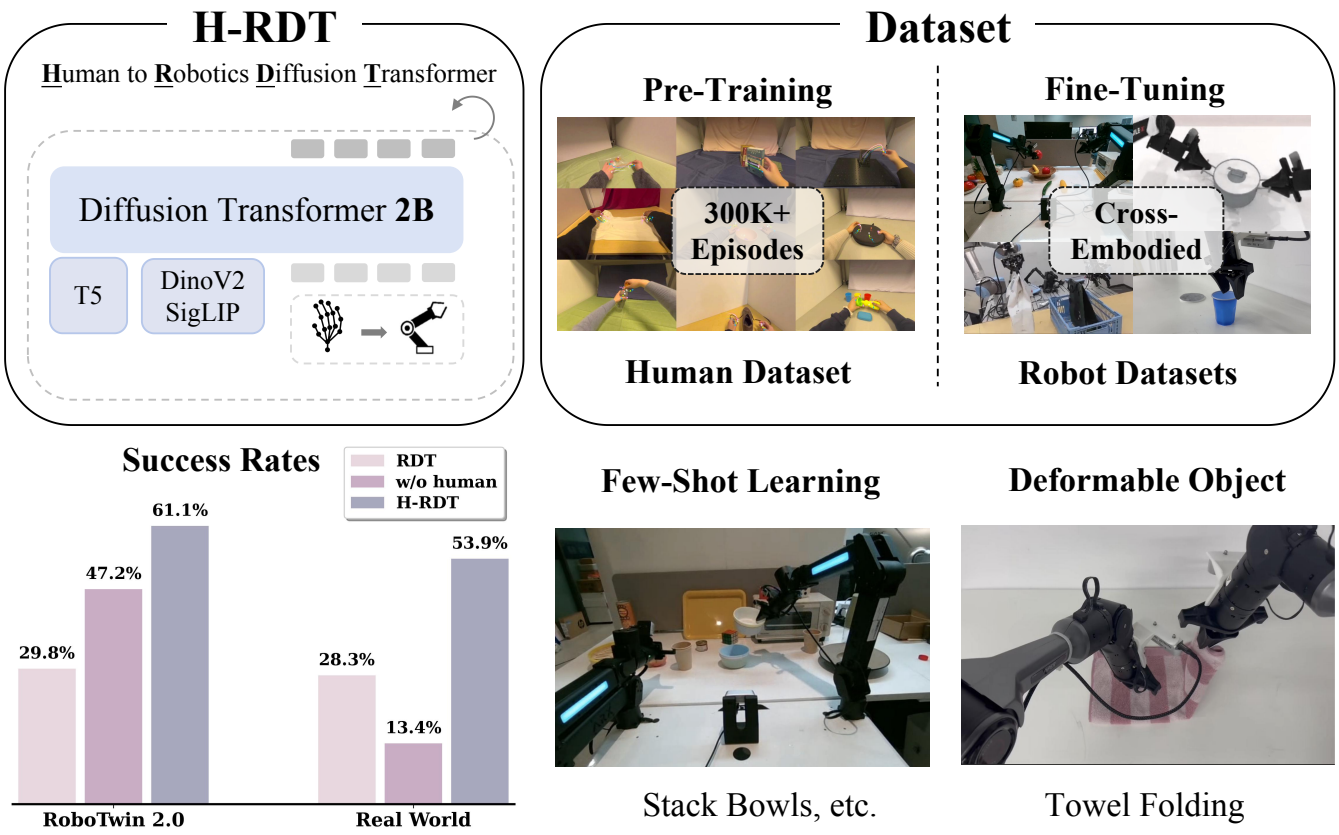


Figure 1: **Overview of H-RDT.** A human-to-robotics diffusion transformer with two-stage training.

task decomposition patterns. **Cross-Embodiment Transfer:** We develop a modular transformer architecture with specialized action encoders and decoders that enable effective knowledge transfer from human demonstrations to diverse robotic platforms while preserving learned manipulation knowledge. **Training Efficiency:** We employ a two-stage training paradigm with flow matching, first pre-training on large-scale human data followed by cross-embodiment fine-tuning, enabling stable and efficient policy learning throughout.

Our method introduces structural and training method innovations for human-to-robot knowledge transfer through human manipulation data pre-training.

The main contributions of this work are:

- A novel framework for systematically leveraging large-scale egocentric human manipulation data to enhance robotic policy learning
- A diffusion transformer architecture with modular human-to-robot transfer components that enables effective cross-embodiment knowledge transfer
- A comprehensive empirical validation demonstrating consistent improvements over state-of-the-art methods across simulation and real-world scenarios
- Insights into the value of human manipulation priors for sample-efficient robot learning, particularly in few-shot settings

## Related Work

### Learning-based Robotic Manipulation

Recent advances in imitation learning have been driven by specialized action policies including ACT (Zhao et al. 2023), Diffusion Policy (Chi et al. 2023), and 3D Diffusion Policy (Ze et al. 2024). These action policies focus on learning direct visuomotor control for manipulation tasks, showing promising results on dexterous manipulation through advanced sequence modeling and generative approaches.

The emergence of Vision-Language-Action (VLA) models represents a significant paradigm shift toward more generalizable robotic systems. Recent VLA approaches include RT-2 (Brohan et al. 2023), OpenVLA (Kim et al. 2024), the Robotics Diffusion Transformer (RDT) (Liu et al. 2024), $\pi_0$ (Black et al. 2024), $\pi_{0.5}$ (Intelligence et al. 2025), and other VLA models (Zhao et al. 2025; Zhen et al. 2024; Liu et al. 2025; Wen et al. 2025b,a). These models combine visual understanding, language comprehension, and action generation within unified architectures, enabling instruction-following capabilities and cross-embodiment generalization through large-scale datasets (O'Neill et al. 2024; Wu et al. 2024; Khazatsky et al. 2024; Fang et al. 2023).

Our work builds upon the RDT architecture while introducing novel structural and training method innovations. Specifically, we adopt flow matching (Lipman et al. 2022; Liu 2022) as our training paradigm, which offers improved

stability and efficiency compared to traditional diffusion training (Esser et al. 2024; Bao et al. 2023). More importantly, we introduce novel human-to-robot knowledge transfer mechanisms that enable large-scale pre-training on human manipulation data followed by cross-embodiment fine-tuning.

### Learning from Egocentric Human Manipulation

Large-scale egocentric datasets (Grauman et al. 2022; Damen et al. 2018; Liu et al. 2022; Banerjee et al. 2025; Grauman et al. 2024) contain tens to hundreds of hours capturing human-object interaction, yet lack precise 3D hand-pose annotations required for dexterous manipulation learning. EgoDex (Hoque et al. 2025) addresses this gap by providing 829 hours (338k episodes) of egocentric video with per-frame 3D hand poses and language descriptions.

EgoMimic (Kareer et al. 2024) and Humanoid Policy (HAT) (Qiu et al. 2025) pioneer the use of egocentric human videos, yet both operate at modest scales: EgoMimic trains on 2k human demos, and HAT on 27k demos—orders of magnitude smaller than the 338k trajectories (829h) employed by H-RDT. Moreover, these works target a single humanoid embodiment; EgoMimic requires paired robot data during co-training, while HAT's retargeting assumes humanoid kinematics. H-RDT, in contrast, decouples large-scale human pre-training from robot-specific fine-tuning and generalizes to arbitrary robot morphologies via modular action adapters. Additional studies explore data augmentation techniques (Li et al. 2025) and paired human–robot data collection (Xie et al. 2025).

## Method

In this section, we present H-RDT (Human to Robotics Diffusion Transformer), a novel approach for leveraging large-scale human manipulation data to enhance robotic policy learning.

### Problem Formulation

We formulate robotic manipulation as a conditional sequence generation problem where the goal is to learn a policy $\pi_\theta$ that generates action sequences $\mathbf{a}_{t:t+H} = \{\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+H-1}\}$ given multimodal observations. Formally, at each timestep $t$, the agent observes visual observations $\mathbf{o}_t \in \mathbb{R}^{H\times W\times 3}$ from one or more RGB cameras, proprioceptive state $\mathbf{s}_t \in \mathbb{R}^{d_s}$ encoding current robot state and gripper state, and language instruction $\mathbf{l} \in \mathbb{R}^{L\times d_{\text{lang}}}$ describing the task. The policy outputs a sequence of future actions $\mathbf{a}_{t:t+H}$ where each action $\mathbf{a}_i \in \mathbb{R}^{d_a}$ represents robot control commands (e.g., joint positions, end-effector poses) over a prediction horizon $H$.

To achieve a generalist policy, large-scale imitation learning is required, while the data for a specified embodiment are scarce. To counter this, current methods majorly turn to train with demonstrations from multiple heterogeneous embodiments (Liu et al. 2024; Black et al. 2024; Intelligence et al. 2025). However, the total scale of the data remains limited due to the high cost of teleoperation.

An alternative approach is to leverage egocentric human manipulation data, which could potentially provide data from a unified human embodiment with manipulation priors transferable across diverse robot platforms, thereby reducing the conflicts of learning from heterogeneous embodiments while enabling low-cost data acquisition. However, this approach faces three main challenges: Firstly, existing methods operate at modest scales with limited human manipulation data, failing to fully exploit the potential of human behavioral priors for robotic learning. Secondly, the significant embodiment differences between humans and robots, including end effector types and forward kinematics (Qiu et al. 2025), make it challenging to effectively transfer manipulation knowledge from human demonstrations to target robots. Thirdly, while concurrent work enables manipulation on specific robots using paired human data (Kareer et al. 2024; Qiu et al. 2025), it remains largely under-addressed how to build a foundation model that can be efficiently adapted to multiple diverse robot embodiments through fine-tuning on robot-specific data.

### Overview

To address the aforementioned challenges, we propose H-RDT (Human to Robotics Diffusion Transformer), as illustrated in Figure 2, a transformer-based architecture trained with a structured paradigm to learn from human data. To counter the embodiment mismatch between humans and robots, H-RDT builds upon a shared action representation space to bridge human and robot embodiments, and satisfies the need for scalable cross-embodiment deployment by utilizing a two-stage training paradigm. Finally, H-RDT leverages flow matching and a scalable Transformer-based architecture for stable and expressive policy learning.

### Human Action Representation Design

To address the challenge of embodiment differences between humans and robots, current methods either use flow as a transit representation for action (Xu et al. 2024; Wen et al. 2024), which provides only high-level object motion guidance without explicit action parameters and requires additional policy networks to translate flow into robot-specific controls, or require detailed re-targeting between human pose and target robot, which constrains the applicability of learning policy. To this end, we utilize detailed 3D hand poses, where actions are represented as compact 48-dimensional vectors capturing essential bimanual dexterous information:

- Bilateral wrist poses (position (3D) and orientation (6D) for both hands): 18 dimensions, which are identical to the End-Effector pose of robots
- Fingertip positions (3D coordinates for all fingers on both hands): 30 dimensions

This representation serves as a superset for the action space of most current robots controlled with the End-Effector poses, thereby ensuring effective knowledge distillation across distinct kinematic structures. This structured encoding explicitly represents fundamental manipulation dynamics and spatial relationships crucial for generalizable

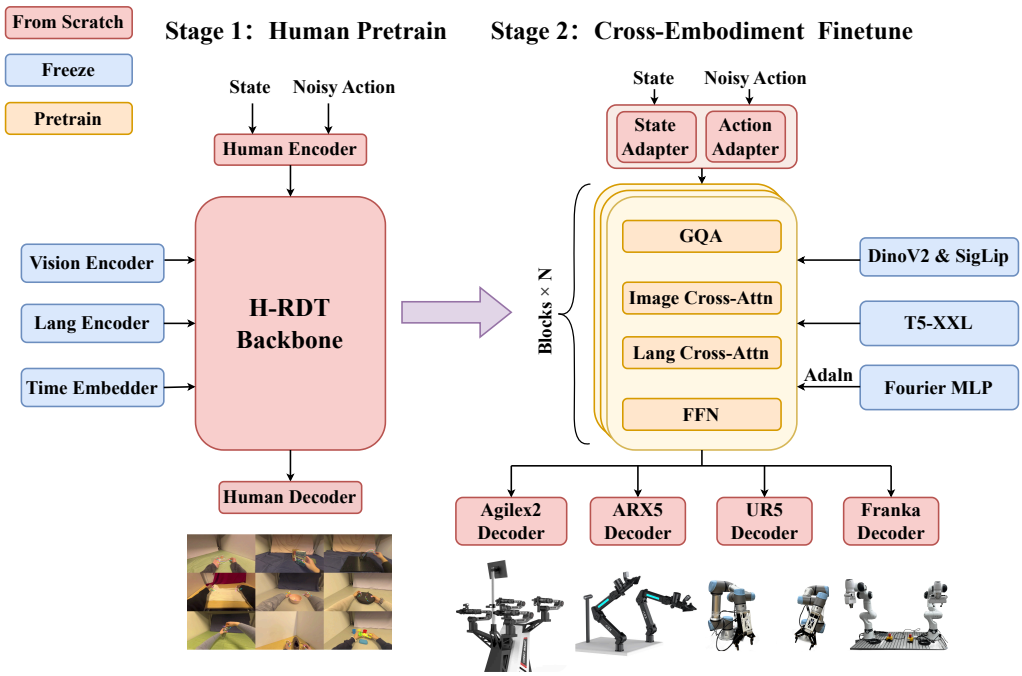


Figure 2: **H-RDT framework.** Our approach consists of two main stages: (1) pre-training on large-scale human manipulation data with 48-dimensional hand pose representations, and (2) cross-embodiment fine-tuning with modular action encoders and decoders adapted to specific robot action spaces.

manipulation, effectively mitigating embodiment discrepancies by focusing on universally transferable features such as grasp configurations, orientation constraints, and relative positional dynamics.

## Two-Stage Training Paradigm

Concurrent work that learns robot policy from human data mostly requires rigorous pairing relationship between human and target embodiment, thereby failing to adapt to multiple embodiments during deployment. To solve this issue, we adopt a carefully designed two-stage training paradigm that maximizes the benefit of human demonstration data while enabling effective cross-embodiment robot deployment. Rather than a traditional diffusion objective, H-RDT employs flow matching (Lipman et al. 2022) for action generation, offering superior training stability and inference efficiency.

**Stage 1: Human Data Pre-training** In the first stage, we train H-RDT on the complete EgoDex dataset with 48-dimensional human hand action representations. Concretely, our model is trained on 338K+ trajectories across 194 distinct manipulation tasks using the complete EgoDex dataset (Hoque et al. 2025), providing comprehensive coverage of human manipulation strategies, object interactions, and bimanual coordination patterns.

**Stage 2: Cross-Embodiment Fine-tuning** To quickly adapt pre-trained for cross-embodiment deployment, the second stage adapts the pre-trained model to specific robot embodiments through selective weight transfer and modular re-initialization: The vision encoder, language encoder, and transformer backbone weights are transferred from the pre-trained model, preserving learned multi-modal representations and manipulation priors developed from human demonstrations. The state adaptor ($\text{MLP}_{\text{state}}$), action adaptor ($\text{MLP}_{\text{action}}$), and action decoder are completely reinitialized to handle the target robot's action space (e.g., 14 dimensions for dual 7-DOF arms with parallel grippers)

This selective transfer strategy ensures that learned manipulation semantics from human demonstrations are preserved while enabling adaptation to diverse robot morphologies. The modular design allows action encoders and decoders to be retrained from scratch for each target embodiment without compromising the learned visual-semantic representations.

## H-RDT Architecture

**Flow Matching for Action Generation** Rather than traditional diffusion training, H-RDT employs flow matching (Lipman et al. 2022) for action generation, offering superior training stability and inference efficiency compared to traditional diffusion modeling. Flow matching learns a vector field that transforms a simple noise distribution to the target action distribution through a continuous normalizing flow.

Given a target action sequence $\mathbf{a}^*_{t:t+H}$, we construct a straight-line flow path:

$$\mathbf{a}_\tau = \tau \cdot \mathbf{a}^*_{t:t+H} + (1-\tau)\mathbf{z} \quad (1)$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise and $\tau \in [0, 1]$ parameterizes the flow time. The neural network $v_\theta$ learns to predict

the vector field:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau,\mathbf{z},\mathbf{a}^*,\mathbf{c}} \left[ \| v_\theta(\mathbf{a}_\tau, \tau, \mathbf{c}) - (\mathbf{z} - \mathbf{a}^*_{t:t+H}) \|^2 \right] \quad (2)$$

where $\mathbf{c} = \{\mathbf{o}_t, \mathbf{s}_t, \mathbf{l}\}$ represents the conditioning information including multi-view RGB observations $\mathbf{o}_t$, proprioception $\mathbf{s}_t$, and language instruction $\mathbf{l}$. During inference, we sample actions by integrating the learned vector field using an ODE solver with deterministic steps (detailed implementation in Appendix B.3).

**Network Architecture** H-RDT adopts a unified transformer architecture comprising five modular components: a vision encoder, language encoder, modular action encoder, transformer backbone, and modular action decoder.

**Vision and Language Encoders:** RGB observations are encoded using pre-trained visual backbones DinoV2 (Oquab et al. 2023) and SigLIP (Zhai et al. 2023), followed by MLP adapters that project image features into the embedding space of dimension $d_{\text{model}}$. Text instructions are embedded using a pre-trained T5-XXL language model (Raffel et al. 2020) and projected via similar adapters.

**Modular Action Encoder:** The proprioceptive state $\mathbf{s}_t$ and noisy action sequence $\mathbf{a}^\tau_{t:t+H}$ are encoded with modular MLP adapters:

$$\mathbf{h}_{\text{state}} = \text{StateAdapter}(\mathbf{s}_t) \in \mathbb{R}^{d_{\text{model}}}, \quad (3)$$

$$\mathbf{h}_{\text{action}} = \text{ActionAdapter}(\mathbf{a}^\tau_{t:t+H}) \in \mathbb{R}^{H \times d_{\text{model}}}, \quad (4)$$

where $\mathbf{a}^\tau_{t:t+H}$ represents the noisy action sequence at flow time $\tau$ used in flow matching training, and $H$ denotes the prediction horizon.

**Transformer Backbone:** H-RDT adopts the LLaMA-3 architecture style (Touvron et al. 2023) with RMSNorm layer normalization and SwiGLU activation functions. Each transformer block processes the concatenated input $\mathbf{x} = \text{Concat}(\mathbf{h}_{\text{state}}, \mathbf{h}_{\text{action}})$ using self-attention, while image and language features are injected via separate cross-attention to avoid modality imbalance (Liu et al. 2024). The flow time $\tau$ is mapped into timestep embeddings and integrated via AdaLN (Peebles and Xie 2023).

**Modular Action Decoder:** The predicted hidden states $\mathbf{h}_{\text{action}}$ are decoded using a modular MLP:

$$\hat{\mathbf{a}}_{t:t+H} = \text{ActionDecoder}(\mathbf{h}_{\text{action}}, \mathbf{t}_{\text{emb}}). \quad (5)$$

where $\mathbf{t}_{\text{emb}}$ represents timestep embeddings for flow matching, and the decoder outputs actions in the target robot's action space. The modular action encoder and decoder are reinitialized during cross-embodiment fine-tuning.

# Experiments

## Experimental Setup

We conduct comprehensive experiments to evaluate H-RDT's effectiveness across simulation and real-world scenarios. Our evaluation covers four key dimensions: (1) single-task and multi-task performance across diverse manipulation scenarios, (2) cross-embodiment generalization across diverse robotic platforms, (3) environmental robustness through domain randomization, and (4) sample efficiency in few-shot learning with limited real-world demonstrations.

**Simulation Environment:** We use the RoboTwin 2.0 platform (Chen et al. 2025), a comprehensive dual-arm manipulation benchmark featuring diverse household tasks. The platform provides two evaluation modes: Easy mode with clean tabletop environments, and Hard mode with domain randomization including 3cm table height variation, random backgrounds, lighting changes, and object clutter.

**Robot Embodiment:** To demonstrate cross-embodiment transfer capabilities, we evaluate H-RDT across multiple robotic platforms in both simulation and real-world settings. Simulation experiments cover two distinct embodiments: Aloha-Agilex-1.0 and dual-arm Franka-Panda. Real-world validation uses three different platforms: dual-arm ARX5, Aloha-Agilex-2.0 (dual-arm Piper), and UR5 + UMI configuration.

## Baselines and Comparison Methods

We compare H-RDT against several state-of-the-art methods:

- **RDT** (Liu et al. 2024): Robotics Diffusion Transformer baseline
- $\boldsymbol{\pi}_0$ (Black et al. 2024): State-of-the-art vision-language-action model
- **w/o human**: Our model without human pre-training

## Real-world Validation

We evaluate H-RDT across three distinct real-world robotic platforms to validate cross-embodiment transfer capabilities and robustness in practical deployment scenarios. All real-world experiments employ multi-task training.

**Aloha-Agilex-2.0 Experiments** We evaluate H-RDT on the Aloha-Agilex-2.0 platform (dual-arm Piper) across two bimanual manipulation tasks.

**Task 1: Towel Folding** This deformable object manipulation task tests the model's ability to handle non-rigid materials through sequential folding operations.

**Task 2: Cup to Coaster Placement** This spatial reasoning task requires the model to automatically select the appropriate hand based on object location: cups on the left side must be grasped with the left hand, while cups on the right side must be grasped with the right hand.

Both tasks employ sub-task-based scoring systems that evaluate progressive completion levels, with final evaluation focusing on complete success rates. Tables 1 and 2 present the performance breakdown across methods, with each task evaluated over 25 trials.

For the towel folding task (Table 1), H-RDT achieves a 52% complete success rate compared to 40% for RDT and 0% for training from scratch. The model without human data fails to achieve any complete folding, showing only partial success at lower skill levels, while RDT and H-RDT demonstrate more sophisticated manipulation capabilities.

The cup-to-coaster task (Table 2) shows H-RDT achieving a 64% complete success rate compared to 28% for RDT and 20% for training without human data. H-RDT demonstrates the lowest failure rate and fewer instances of partial

| Performance Level | RDT | w/o human | H-RDT |
|---|---|---|---|
| 0.0: Complete failure | - | 7 | - |
| 0.2: One fold single side | - | 18 | - |
| 0.5: One fold both sides | 3 | - | 3 |
| 0.7: Two fold incomplete | 12 | - | 9 |
| 1.0: Two fold complete | 10 | - | 13 |
| Complete success rate | 40% | 0% | **52%** |

Table 1: Towel folding task results on Aloha-Agilex-2.0 platform with detailed performance breakdown.

| Performance Level | RDT | w/o human | H-RDT |
|---|---|---|---|
| 0.0: Complete failure | 10 | 14 | 6 |
| 0.4: Grasped, failed to place | 8 | 6 | 3 |
| 1.0: Successfully completed | 7 | 5 | 16 |
| Complete success rate | 28% | 20% | **64%** |

Table 2: Cup to coaster placement task results on Aloha-Agilex-2.0 platform with detailed performance breakdown.

success, indicating more robust performance for tasks requiring spatial reasoning to select appropriate arms.

Overall, H-RDT achieves an average success rate of 58% across both bimanual tasks compared to 34% for RDT and 10% for training without human data, demonstrating the effectiveness of human manipulation priors for diverse coordination challenges including deformable object manipulation and spatial reasoning tasks.

**Dual-arm ARX5 Few-shot Experiments** To thoroughly validate the advantages of human manipulation priors, we design a challenging real-world experiment with a combination of massive task diversity and data scarcity: 113 diverse pick-and-place tasks using a dual-arm ARX5 robotic system, with only 1-5 demonstrations per task. This multi-task few-shot setting is specifically designed to test the limits of sample efficiency and highlight the value of human behavioral priors.

The EgoDex pretraining dataset contains extensive pick-and-place manipulation patterns similar to these tasks, providing rich prior knowledge about how to perform such operations. Under these challenging conditions—where even state-of-the-art models like $\pi_0$ struggle to properly fit the limited demonstration trajectories—H-RDT's human manipulation priors enable noticeable performance improvements. H-RDT achieves an average success rate of 41.6% compared to 16.0% for RDT, 31.2% for $\pi_0$, and 17.6% for H-RDT w/o human, demonstrating the value of human manipulation priors for few-shot learning in data-limited scenarios.

**Dual UR5 + UMI Experiments** We evaluate H-RDT on a dual UR5 robotic system with demonstrations collected using Universal Manipulation Interface (UMI) (Chi et al. 2024), a data collection framework that enables portable, low-cost human demonstration collection through hand-held grippers.

The evaluation focuses on bimanual takeout bag placement tasks decomposed into four sequential subtasks: right-hand pick, right-hand place, left-hand pick, and left-hand place.

Table 4 presents the success rates for each subtask across different methods, with each evaluation conducted over 25 trials.

H-RDT achieves consistently superior performance across all subtasks, with an average success rate of 58.0% compared to 29.0% for RDT, 31.0% for $\pi_0$, and 16.0% for training from scratch. The results show notable improvements in pick operations (64% and 60% for right and left hands, respectively) and 27-42% absolute improvements over baseline methods, demonstrating the value of human manipulation priors for bimanual coordination.

## Simulation Results on RoboTwin 2.0

**Single-Task Performance** We evaluate single-task performance on 13 representative manipulation tasks from the RoboTwin 2.0 benchmark. Each task is trained on 50 demonstrations collected in clean environments and evaluated in two modes: Easy mode (clean tabletop environments matching training conditions) and Hard mode (challenging environments with domain randomization including lighting changes, object clutter, and table height variations). Detailed results for all tasks are provided in Table 8 in the Appendix.

H-RDT achieves the highest average success rate of 68.7% in Easy mode and 25.6% in Hard mode, demonstrating significant improvements over existing methods. H-RDT substantially surpasses training from scratch (w/o human) by 8.4% in both Easy and Hard modes, validating the effectiveness of human manipulation pre-training.

**Multi-Task Performance** We conduct multi-task experiments on 45 tasks from RoboTwin 2.0, training on approximately 2250 demonstrations collected under domain randomization (Hard mode data). Table 5 shows the results on a representative subset of 10 tasks evaluated in Hard mode.

In multi-task settings, H-RDT achieves an average success rate of 87.2%, significantly outperforming RDT (28.8%), $\pi_0$ (48.4%), and H-RDT w/o human (67.2%). H-RDT shows a substantial 20.0% absolute improvement over training from scratch, which is notably larger than the improvements observed in single-task scenarios, demonstrating that human manipulation pre-training provides even greater advantages when learning across diverse tasks simultaneously.

**Cross-Embodiment Generalization** To further validate the cross-embodiment transfer capabilities of H-RDT, we conduct multi-task experiments across two different robotic embodiments in simulation. We evaluate both Aloha-Agilex-1.0 and Franka-Panda platforms using the same experimental setup as described above. Figure 3 shows the performance comparison across these platforms.

H-RDT demonstrates strong performance across both embodiments, achieving 87.2% on Aloha-Agilex-1.0 and 62.9% on Franka-Panda, significantly outperforming baseline methods on both platforms. Detailed per-task results for Franka-Panda are provided in Table 6 in the Appendix. The consistent improvements across different robotic morphologies validate the cross-embodiment generalization capabilities of our modular action encoder design.

| Task Category | RDT | $\pi_0$ | w/o human | H-RDT |
|---|---|---|---|---|
| Pick yellow item to the plate | 12% | 16% | 28% | **40%** |
| Place banana and carrot in basket | 24% | 40% | 8% | **44%** |
| Stack two bowls together | 32% | 60% | 40% | **68%** |
| Place cube in front of highest chips | 12% | 0% | 0% | **20%** |
| Stack one can on top of the other can | 0% | **40%** | 12% | 36% |
| Average | 16.0% | 31.2% | 17.6% | **41.6%** |

Table 3: Real-world few-shot learning results on manipulation tasks. Each task has 1-5 demonstrations for training. H-RDT demonstrates superior sample efficiency and real-world transfer capabilities.

| Subtask | RDT | $\pi_0$ | w/o human | H-RDT |
|---|---|---|---|---|
| Right hand pick | 36% | 40% | 20% | **64%** |
| Right hand place | 32% | 28% | 16% | **56%** |
| Left hand pick | 28% | 36% | 20% | **60%** |
| Left hand place | 20% | 20% | 8% | **52%** |
| Average | 29.0% | 31.0% | 16.0% | **58.0%** |

Table 4: Performance breakdown on dual UR5 + UMI takeout bag placement task showing success rates for individual subtasks.

| Task | RDT | $\pi_0$ | w/o human | H-RDT |
|---|---|---|---|---|
| Click Alarmclock | 30% | 69% | 69% | **94%** |
| Click Bell | 13% | 40% | 77% | **98%** |
| Dump Bin Bigbin | 22% | 33% | 58% | **89%** |
| Grab Roller | 57% | 71% | 91% | **97%** |
| Handover Mic | 81% | 97% | 93% | **99%** |
| Move Playingcard | 3% | 23% | 44% | **67%** |
| Open Laptop | 33% | 35% | 88% | **92%** |
| Open Microwave | 19% | 46% | 32% | **82%** |
| Press Stapler | 15% | 59% | 71% | **81%** |
| Stack Bowls Three | 15% | 11% | 49% | **73%** |
| Average | 28.8% | 48.4% | 67.2% | **87.2%** |

Table 5: Multi-task success rates (%) on RoboTwin 2.0 benchmark (Hard mode evaluation). Results show performance after multi-task training on 50 tasks with domain randomized data.

## Analysis and Discussion

**Impact of Human Pre-training:** The consistent improvements of H-RDT over the baseline without human data across all experimental settings validate our core hypothesis that human manipulation data provides valuable inductive biases. The benefits are most pronounced in few-shot real-world scenarios, where human priors about object affordances and manipulation strategies prove crucial.

**Environmental Robustness:** H-RDT demonstrates strong performance under challenging conditions in RoboTwin 2.0 Hard mode with domain randomization. The model successfully handles environmental variations including lighting changes, object clutter, and table height variations, consistently outperforming baseline methods.

**Sample Efficiency:** In few-shot real-world experiments, H-RDT's ability to learn from just 1-5 demonstrations per task significantly outperforms baselines, highlighting the practical value of human behavior priors for reducing data requirements in robotic learning.

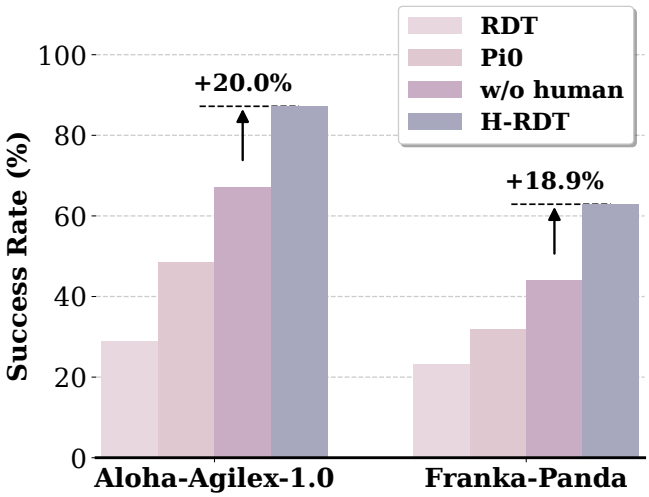


Figure 3: Cross-embodiment multi-task performance on RoboTwin 2.0 tasks.

**Task Diversity and Complexity:** Real-world experiments demonstrate H-RDT's capability to handle diverse manipulation challenges, including deformable object manipulation and tasks requiring spatial reasoning, showcasing its versatility across different manipulation complexities.

**Cross-Platform Robustness:** Our comprehensive evaluation across both simulation and real-world settings demonstrates H-RDT's robust performance across diverse robotic embodiments, including Aloha-Agilex-1.0, dual-arm Piper, dual-arm ARX5, dual-arm Franka-Panda, and dual-arm UR5+UMI platforms. This cross-platform consistency validates the effectiveness of our modular architecture design and human-to-robot knowledge transfer approach.

## Conclusion

This paper introduces H-RDT, a novel approach that leverages large-scale egocentric human manipulation videos with 3D hand pose annotations to enhance robotic manipulation capabilities. Our central contribution shows that rich manipulation knowledge can be learned from human behavioral priors and adapted to diverse robotic manipulation tasks.

Comprehensive evaluations demonstrate consistent improvements over state-of-the-art methods, validating that human manipulation priors provide powerful inductive biases for sample-efficient robotic learning.

## References


Aldaco, J.; Armstrong, T.; Baruch, R.; Bingham, J.; Chan, S.; Draper, K.; Dwibedi, D.; Finn, C.; Florence, P.; Goodrich, S.; et al. 2024. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*.

Banerjee, P.; Shkodrani, S.; Moulon, P.; Hampali, S.; Han, S.; Zhang, F.; Zhang, L.; Fountain, J.; Miller, E.; Basol, S.; et al. 2025. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7061–7071.

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22669–22679.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Bu, Q.; Cai, J.; Chen, L.; Cui, X.; Ding, Y.; Feng, S.; Gao, S.; He, X.; Hu, X.; Huang, X.; et al. 2025. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*.

Chen, T.; Chen, Z.; Chen, B.; Cai, Z.; Liu, Y.; Liang, Q.; Li, Z.; Lin, X.; Ge, Y.; Gu, Z.; et al. 2025. RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation. *arXiv preprint arXiv:2506.18088*.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

Chi, C.; Xu, Z.; Pan, C.; Cousineau, E.; Burchfiel, B.; Feng, S.; Tedrake, R.; and Song, S. 2024. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, 720–736.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Fang, H.-S.; Fang, H.; Tang, Z.; Liu, J.; Wang, C.; Wang, J.; Zhu, H.; and Lu, C. 2023. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18995–19012.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Hoque, R.; Huang, P.; Yoon, D. J.; Sivapurapu, M.; and Zhang, J. 2025. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. *arXiv preprint arXiv:2505.11709*.

Intelligence, P.; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; et al. 2025. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*.

Kareer, S.; Patel, D.; Punamiya, R.; Mathur, P.; Cheng, S.; Wang, C.; Hoffman, J.; and Xu, D. 2024. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*.

Khazatsky, A.; Pertsch, K.; Nair, S.; Balakrishna, A.; Dasari, S.; Karamcheti, S.; Nasiriany, S.; Srirama, M. K.; Chen, L. Y.; Ellis, K.; et al. 2024. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Li, Z.; Wang, J.; Chen, H.; Liu, M.; et al. 2025. H2R: Learning Human-to-Robot Imitation with Data Augmentation. *arXiv preprint arXiv:2501.xxxxx*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, J.; Chen, H.; An, P.; Liu, Z.; Zhang, R.; Gu, C.; Li, X.; Guo, Z.; Chen, S.; Liu, M.; et al. 2025. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*.

Liu, Q. 2022. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*.

Liu, S.; Wu, L.; Li, B.; Tan, H.; Chen, H.; Wang, Z.; Xu, K.; Su, H.; and Zhu, J. 2024. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*.

Liu, Y.; Liu, Y.; Jiang, C.; Lyu, K.; Wan, W.; Shen, H.; Liang, B.; Fu, Z.; Wang, H.; and Yi, L. 2022. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21013–21022.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

O'Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 6892–6903. IEEE.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Qiu, R.-Z.; Yang, S.; Cheng, X.; Chawla, C.; Li, J.; He, T.; Yan, G.; Yoon, D. J.; Hoque, R.; Paulsen, L.; et al. 2025. Humanoid Policy˜ Human Policy. *arXiv preprint arXiv:2503.13441*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Team, O. M.; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Kreiman, T.; Xu, C.; et al. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Wen, C.; Lin, X.; Wei, J. S.; Chen, K.; and Gao, Y. 2024. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*.

Wen, J.; Zhu, Y.; Li, J.; Tang, Z.; Shen, C.; and Feng, F. 2025a. DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control. *arXiv preprint arXiv:2502.05855*.

Wen, J.; Zhu, Y.; Li, J.; Zhu, M.; Tang, Z.; Wu, K.; Xu, Z.; Liu, N.; Cheng, R.; Shen, C.; et al. 2025b. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*.

Wu, K.; Hou, C.; Liu, J.; Che, Z.; Ju, X.; Yang, Z.; Li, M.; Zhao, Y.; Xu, Z.; Yang, G.; et al. 2024. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*.

Xie, S.; Cao, H.; Weng, Z.; Xing, Z.; Shen, S.; Leng, J.; Qiu, X.; Fu, Y.; Wu, Z.; and Jiang, Y.-G. 2025. Human2Robot: Learning Robot Actions from Paired Human-Robot Videos. *arXiv preprint arXiv:2502.16587*.

Xu, M.; Zhang, H.; Hou, Y.; Xu, Z.; Fan, L.; Veloso, M.; and Song, S. 2025. DexUMI: Using Human Hand as the Universal Manipulation Interface for Dexterous Manipulation. *arXiv preprint arXiv:2505.21864*.

Xu, M.; Zhang, Z.; Chi, C.; and Song, S. 2024. Flow as the Cross-Domain Manipulation Interface. *arXiv preprint arXiv:2407.15208*.

Ze, Y.; Zhang, G.; Zhang, K.; Hu, C.; Wang, M.; and Xu, H. 2024. 3d diffusion policy. *arXiv e-prints*, arXiv–2403.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1702–1713.

Zhao, T. Z.; Kumar, V.; Levine, S.; and Finn, C. 2023. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*.

Zhen, H.; Qiu, X.; Chen, P.; Yang, J.; Yan, X.; Du, Y.; Hong, Y.; and Gan, C. 2024. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*.