

Towards Scalable Robot Learning: From Teleoperation to Web-scale Data

Chaitanya Chawla

CMU-RI-TR-26-72

July 10, 2026

School of Computer Science
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Thesis Committee:

Prof. Guanya Shi, Chair

Prof. Laszlo Jeni

Eliot Xing

*Submitted in partial fulfillment of the
requirements for the degree of Master of Science in Robotics*

Dedicated to my family

Acknowledgments

I am deeply grateful to my advisor, Guanya, for thoughtful mentorship and for creating the supportive environment of the LeCAR Lab, where I have grown immensely. Your guidance taught me to ask the right questions and approach problems from first principles.

I am also grateful to the Bot Intelligence Group for its support and encouragement, which helped make my path to CMU possible. In particular, I thank Jean, Tanmay, who advised my first project at CMU, Eliot, Peter, and Zhixuan for their mentorship, guidance, and belief in me.

To my labmates—Yuanhang, Tairan, Chaoyi, Zeji, Yifu, Nikhil, and Yogita—thank you for your generous help, stimulating conversations, and all the fun times along the way. You made this journey both more productive and more memorable.

I also thank the Art of Living family in Pittsburgh, especially Gaurav, for giving me a home away from home.

I am especially thankful to my parents and grandparents for their unwavering love and support from afar. Finally, thank you to Shiqi for always being there; your friendship has meant more to me than I can ever express.

Abstract

Humanoid robots operating in human environments must manipulate articulated objects under contact and kinematic constraints that human demonstrations do not satisfy. That mismatch makes the human–humanoid embodiment gap the central bottleneck for learning from human data: robot demonstrations are expensive and sparse, while human demonstrations inhabit a different state-action space and often violate robot kinematic constraints. This thesis studies how to convert human behavior into supervision that remains executable for the target robot body.

The first part develops **HAT (Humanoid Policy $\sim$ Human Policy)** for cross-embodiment supervision in humanoid manipulation. It places humans and humanoids in a unified state-action representation, enabling a transformer policy to co-train on human and robot demonstrations and retarget its predictions at deployment. To support this formulation, we introduce **PH²D**, a task-oriented egocentric human demonstration dataset that expands data scale without discarding embodiment structure.

The second part presents **EmbodimentHOI**, which addresses a harder embodiment-gap setting in dexterous hand-object interaction. It starts from a flow-matching diffusion model trained in human hand-object space, then applies a differentiable guidance function during sampling to steer trajectories toward a target humanoid embodiment, jointly optimizing wrist reachability and base placement before downstream control.

Together, these chapters show that scalable robot manipulation requires data transformations that preserve task structure while respecting the robot body.

Table of Contents

List of Figures	viii
List of Tables	x
1 Introduction	1
2 HAT: Humanoid Policy $\sim$ Human Policy	3
2.1 Introduction	3
2.2 Related Work	4
2.3 PH ² D: Task-Oriented Human-Humanoid Data	5
2.4 Human Action Transformer	6
2.5 Experiments	8
2.6 Conclusion	12
3 EmbodyHOI: Embodiment-Aware Generation of Hand–Object Interactions	15
3.1 Introduction	15
3.2 Related Work	17
3.3 Method	17
3.4 Experimental Results	20

3.5	Conclusion	23
4	Conclusion and Future Work	26
4.1	Future Work	27
	Bibliography	29
	Appendix A Humanoid Policy Appendix	43
A.1	More Ablation Study - Data Collection	43
A.2	In-Depth Analysis of Different Types of Generalization	44
A.3	In-Depth Comparison between Humanoid A and Humanoid B Configurations .	47
	Appendix B EmbodyHOI Project Appendix	49
B.1	Flow Matching and Energy Guidance	49
B.2	Policy Observations	49
B.3	IK Optimization Details	50
B.4	Additional Singularity Analysis on Unitree H1-2	51
B.5	Reward and Termination Design	51
B.6	Training Hyperparameters	52
B.7	Computational Cost	52
B.8	Additional Results	53

List of Figures

Figure 2.1	PH ² D data collection setup	6
Figure 2.2	Overview of HAT	7
Figure 2.3	Humanoid platforms used in HAT experiments	9
Figure 2.4	Few-shot transfer across objects	12
Figure 2.5	Few-shot transfer as Humanoid B demonstrations increase	13
Figure 3.1	EMBODYHOI pipeline	18
Figure 3.2	Kinematic conditioning for Unitree G1	21
Figure 3.3	Reward design ablation	24
Figure 3.4	Qualitative cross-embodiment adaptation for Unitree H1-2 and G1	25
Figure A.1	Cup passing across four backgrounds	43
Figure A.2	Horizontal grasping across four objects	43
Figure A.3	Vertical grasping across nine positions	44
Figure A.4	Pouring across robot-table configurations	44
Figure A.5	Object placement generalization	46
Figure B.1	Kinematic conditioning for Unitree H1-2	52
Figure B.2	Additional qualitative manipulation tasks on Unitree G1	54
Figure B.3	Unguided vs. IK-guided MAD-HOI generations	56

List of Tables

Table 2.1	HAT task-oriented egocentric dataset scale	5
Table 2.2	HAT data collection efficiency	5
Table 2.3	Autonomous execution success on Humanoid A	9
Table 2.4	Background generalization in cup passing	10
Table 2.5	Object appearance generalization in horizontal grasping	10
Table 2.6	Ablation on unified state space and action speed alignment	14
Table 3.1	Generation quality on 419 ARCTIC examples	21
Table 3.2	IK-guided sampling ablation	22
Table 3.3	End-to-end G1 reference-tracking success from guided vs. unguided samples	23
Table A.1	Whole-body movement ablation	45
Table A.2	Object appearance generalization in horizontal grasping	47
Table A.3	Background generalization in cup passing	47
Table A.4	Joint range of motion comparison	48
Table B.1	Policy observations for articulated-object manipulation	50
Table B.2	IK guidance optimization objective	51
Table B.3	IK guidance optimization parameters	52
Table B.4	Reward terms for task execution and stabilization	53

Table B.5	Episode termination conditions during training	53
Table B.6	Embodiment RL hyperparameters	55
Table B.7	Approximate runtime of major pipeline stages	55

Chapter 1

Introduction

General-purpose robot manipulation requires policies that transfer across objects, environments, tasks, and embodiments. This remains difficult because manipulation is not only a perception problem but an embodied interaction problem: success depends on contact, reachability, hand morphology, viewpoint, and control constraints. A policy that succeeds for one object placement, camera view, or robot body can fail under small changes to geometry or embodiment.

Recent progress in robot learning has come from imitation learning, teleoperation, and large visuomotor policies. But unlike language or vision, robot manipulation cannot be scaled simply by collecting more generic data. The most useful supervision must encode how bodies move, contact objects, and recover from physical errors. Such data is expensive to collect on real robots, requires hardware, resets, and operator time, and is usually tied to a specific embodiment. As a result, current manipulation policies are often trained on narrow datasets and struggle to generalize beyond the settings seen during collection.

This bottleneck is especially severe for humanoid and dexterous manipulation. These systems must coordinate whole-arm reaching, wrist orientation, multi-finger contact, and object motion under tight kinematic limits. Small errors in pose, scale, or timing can break the interaction entirely. Scaling such skills with robot-only data is therefore doubly difficult: the tasks are more contact-sensitive, and the hardware needed to collect large datasets is more complex and expensive. A scalable solution must reduce dependence on robot-collected demonstrations without discarding the embodiment constraints that determine whether a behavior is executable.

This thesis studies a central question: how can robot learning scale beyond robot-only data collection without losing the embodiment-specific structure required for execution? Human behavior is a natural answer because humans perform rich manipulation across diverse settings at massive scale. However, raw human data is not directly usable for robots. Human hands, robot hands, and humanoid arms differ in kinematics, dexterity, and reachability; internet videos lack action labels and calibrated geometry; and trajectories that are valid for

humans may be infeasible for robots. The core challenge is therefore not only to access more data, but to transform abundant human behavior into training signals that remain faithful to robot embodiment.

This thesis argues that scalable manipulation requires both data scale and data alignment. More data helps only when the representation preserves the structure that a robot must ultimately execute: where the hands move, how contact evolves, and whether the target embodiment can realize the behavior. This perspective motivates two complementary approaches. One is to align human demonstrations with robot demonstrations in a shared representation for direct policy learning. The other is to generate interaction trajectories from human data while incorporating embodiment constraints before downstream control. This thesis addresses this question through two projects on scalable robot learning for manipulation. Chapter 2 presents **HAT (Humanoid Policy $\sim$ Human Policy)**, which uses task-oriented egocentric human demonstrations as cross-embodiment supervision for humanoid manipulation. HAT places humans and humanoids in a unified state-action representation, enabling co-training with human and robot demonstrations and retargeting to robot actions at deployment. Chapter 3 presents **EmbodimentHOI**, which extends this direction to dexterous hand-object interaction. It guides text-conditioned hand-object motion generation with differentiable inverse kinematics on the target humanoid, producing interaction references that better respect robot reachability and dexterous hand constraints before downstream policy learning.

Together, these projects study two complementary solutions to the same bottleneck: how to scale manipulation learning beyond robot-only data collection while accounting for the embodiment gap. HAT focuses on learning directly from aligned human demonstrations, while EmbodimentHOI focuses on generating embodiment-aware interaction trajectories for downstream control. Across both projects, the central claim is that scalable robot manipulation requires not only more data, but data that is aligned to the structure and constraints of the robot body.

Chapter 2

HAT: Humanoid Policy $\sim$ Human Policy

This chapter studies the thesis problem from the perspective of direct policy learning: can humanoid manipulation scale beyond robot-only data collection if human demonstrations are aligned tightly enough to robot execution? HAT answers this question with a policy learning framework that uses egocentric human demonstrations as scalable cross-embodiment data for humanoid manipulation.

2.1 Introduction

Learning manipulation policies directly from robot demonstrations has enabled strong progress in bimanual manipulation, visuomotor control, and dexterous teleoperation. However, real robot data remains expensive to collect. Each demonstration requires a physical platform, safe teleoperation, scene resets, and platform-specific hardware. This cost is especially limiting for humanoid robots, where whole-body posture, active head motion, bimanual reaching, and dexterous hands make data collection slower than ordinary human manipulation.

Within the broader thesis argument developed in Chapter 1, HAT focuses on the most direct route to scalability: learning from human demonstrations themselves rather than from generated motion priors or robot-only data. Human demonstrations provide a natural alternative source of manipulation experience. Humans already perform the contact-rich skills that humanoid robots aim to learn: grasping, passing, pouring, and placing objects under diverse appearances, backgrounds, and spatial configurations. The challenge is that human behavior is not directly robot data. Human and humanoid hands differ in morphology, humans move faster than teleoperated robots, and the visual observations collected from wearable devices differ from the robot’s egocentric camera stream.

HAT addresses this problem by treating humans and humanoid robots as different embodiments in a shared policy space. The key idea is to collect task-oriented egocentric human

demonstrations with accurate head, wrist, and hand pose supervision, and to train a transformer policy in a unified human-humanoid state-action representation. During training, the policy sees both human and robot demonstrations. During deployment, its predictions are retargeted to robot wrist and finger commands. The practical result is a policy that learns from larger and more diverse human data while still executing in robot joint space.

The resulting system shows that human demonstrations can improve humanoid manipulation without requiring a robot for every data collection session. Co-training with human data preserves in-distribution performance while improving robustness to backgrounds, object appearances, and object placements that are absent from the robot demonstrations but present in the human data.

At the thesis level, this chapter addresses the regime in which human behavior can be brought close enough to robot execution through representation alignment and retargeting. It does not yet solve the harder setting where fine-grained contact and dexterous embodiment mismatch dominate execution. That remaining gap motivates the next chapter, which studies how embodiment constraints can be injected earlier, during motion generation itself.

2.2 Related Work

Robot manipulation from demonstrations. Behavior cloning, action chunking, and large-scale robot data collection have made demonstration learning a practical route to visuomotor control [1–6]. These methods work well when robot data are plentiful, but they remain anchored to the robot embodiment that produced the demonstrations. HAT follows the same imitation-learning objective while asking whether a smaller robot dataset can be strengthened with a much larger source of human demonstrations.

Human data for manipulation learning. Human videos contain a broad range of objects, tasks, and scene conditions, making them a natural source of manipulation priors. Prior work has used human video to learn affordances, object-centric structure, and dexterous skills [7–54]. These methods often rely on intermediate perception models or weak supervision. HAT instead uses egocentric human demonstrations with explicit hand and head pose annotations, so the human data can be consumed directly by the policy.

Cross-embodiment policies. Cross-embodiment learning studies how to share data across agents with different kinematics, sensors, and action spaces [55, 55, 56, 56]. For humanoids, prior teleoperation systems retarget human motion to robot actions, and human-shadowing approaches use robots to imitate human motion online [55–57, 57, 58, 58]. HAT keeps the retargeting stage, but removes the requirement that every useful demonstration be collected through the robot itself.

Table 2.1. Scale of task-oriented egocentric datasets used in HAT. PH²D contains both human and humanoid demonstrations, with human demonstrations providing the larger and more scalable data source.

Dataset	Human Frames	Human Demos	Robot Frames	Robot Demos
DexCap	~378k	787	N/A	N/A
EgoMimic	~432k	2,150	~1.29M	1,000
PH ² D	~3.02M	26,824	~668k	1,552

Table 2.2. Amortized mean and standard deviation of time per demonstration, including scene resets. Wearing the VR capture device has little effect on human collection time, while humanoid VR teleoperation is substantially slower.

Collection mode	Grasping (s)	Pouring (s)
Human demonstration	3.79 ± 0.27	4.81 ± 0.35
Human demonstration with VR	4.09 ± 0.30	4.90 ± 0.26
Humanoid demonstration (VR teleoperation)	19.72 ± 1.65	37.31 ± 6.25

Task-oriented egocentric data. Egocentric datasets are useful only when their tasks and action supervision match the target robot problem. Systems such as DexCap and EgoMimic collect manipulation demonstrations with hand-pose or visual alignment [59, 59, 60, 60, 61]. HAT builds on this direction with PH²D, a task-oriented human-humanoid dataset designed around manipulation tasks that directly transfer to humanoid execution.

2.3 PH²D: Task-Oriented Human-Humanoid Data

PH²D is designed around a simple requirement: human demonstrations should match the manipulation tasks that the humanoid policy must execute. Instead of collecting unconstrained daily activity, operators wear consumer VR devices and perform task-oriented manipulation such as grasping, passing, and pouring. Each demonstration contains egocentric visual observations, language instructions, head pose, wrist pose, and hand keypoints.

Table [Table 2.1](#) summarizes the scale of PH²D relative to other task-oriented egocentric datasets. Data collection uses two hardware configurations. The first uses the Apple Vision Pro with its built-in camera and ARKit hand tracking. The second uses either a Meta Quest 3 or Apple Vision Pro with a mounted ZED stereo camera, using OpenTeleVision for head and hand pose capture [61]. These devices provide 3D head and hand pose estimates while keeping the setup portable and less expensive than collecting every demonstration through a humanoid robot. Table [Table 2.2](#) quantifies this advantage. VR collection takes 4.09 s for grasping and 4.90 s for pouring, compared with 19.72 s and 37.31 s for humanoid



Figure 2.1. PH²D data collection setup using consumer VR devices and egocentric cameras. The setup records visual observations together with tracked head, wrist, and hand keypoints, allowing human demonstrations to be aligned with humanoid manipulation.

VR teleoperation; the uninstrumented human timings show that wearing the capture device adds little overhead. Appendix A contains the human-data ablation, the background and appearance breakdowns, and the arm-range comparison between the two humanoids.

Figure [Figure 2.1](#) should be placed near the dataset description because the data collection hardware motivates the scalability of the method. The important design choice is not only the use of wearable devices, but the alignment between human and robot tasks. Operators perform manipulation skills that overlap with robot execution, and each sequence is synchronized across visual and proprioceptive streams.

Human demonstrations introduce two action-domain gaps. First, humans naturally move their torso, shoulders, and waist during manipulation, while the evaluated humanoid policies focus on head, arm, wrist, and hand control. PH²D reduces this gap by asking human demonstrators to sit upright and minimize whole-body motion. Second, humans execute manipulation faster than humanoid teleoperation. HAT slows human trajectories during training by interpolating translations and rotations, using a slowdown factor of $\alpha_{\text{slow}} = 4$ across tasks.

These choices matter because the model learns from mixed human and robot data in a single transformer. Without explicit alignment in time and representation, the data sources would remain distinguishable and the benefit of human data would be much smaller.

2.4 Human Action Transformer

The method is therefore designed around a single thesis-level requirement: preserve as much scalable human variation as possible while keeping the learned representation executable on a humanoid body.

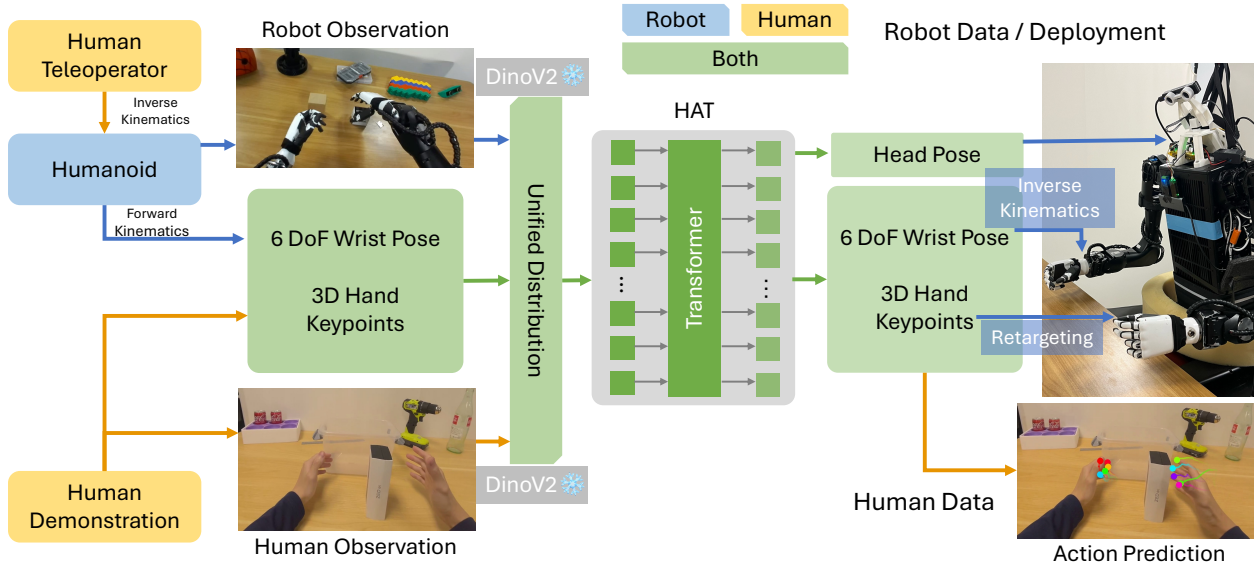


Figure 2.2. Overview of HAT. The policy samples human and robot demonstrations during training, encodes egocentric images with a frozen DINOv2 backbone, predicts wrist and fingertip trajectories in a unified human-centric action space, and retargets the prediction to robot actions for deployment.

HAT learns a policy from a mixture of robot demonstrations and human demonstrations. Let $D_{\text{robot}} = \{(S_i, A_i)\}_{i=1}^N$ denote teleoperated humanoid demonstrations, where S_i contains visual and proprioceptive observations and A_i contains robot actions. Let $D_{\text{human}} = \{(\tilde{S}_i, \tilde{A}_i)\}_{i=1}^M$ denote the PH²D human dataset. Since human demonstrations are substantially easier to collect, the intended regime is $M \gg N$.

The goal is to learn a policy $\pi : S \rightarrow A$ that predicts future robot action chunks from the current egocentric observation. HAT follows an action-chunking transformer architecture, but changes the state and action representation so that both humans and humanoids can be modeled by the same network.

The full HAT pipeline is summarized in Figure [Figure 2.2](#). Images are encoded with a frozen DINOv2 ViT-S backbone [62]. Proprioception is represented in a human-centric space rather than raw robot joints. The observation vector contains 6D rotations for the head, left wrist, and right wrist; 3D positions for the left and right wrists; and 3D positions for ten fingertips. This produces a 54-dimensional proprioceptive representation shared by humans and humanoids.

The shared representation is possible because both the human demonstrator and the robot are modeled through end-effectors and fingertips. For the humanoids used in the HAT experiments, the robot hands have five fingers, enabling a direct fingertip correspondence

with human hands. The same principle could also support simpler grippers by mapping a subset or function of human fingertip motion to the robot action space.

At deployment, HAT predictions are converted into robot commands through inverse kinematics and hand retargeting. This separates the policy’s learning problem from the low-level joint representation while preserving embodiment constraints at execution time. The policy is trained with an ℓ_1 action prediction loss and an additional end-effector translation term:

$$\mathcal{L} = \ell_1(\pi(s_i), a_i) + \lambda \ell_1(\pi(s_i)_{\text{EEF}}, a_{i,\text{EEF}}),$$

where EEF denotes the left and right wrist translation dimensions and $\lambda = 2$ emphasizes accurate end-effector motion.

This loss is intentionally simple. The point is not to learn a full generative world model, but to exploit a representation that is shared across embodiments and sufficiently structured to support retargeting. In practice, HAT uses a frozen DINOv2 visual backbone [62] together with an action-chunking transformer, so the technical burden is concentrated on the state-action alignment rather than on the visual encoder.

The method also addresses visual-domain gaps. Human demonstrations may use cameras that differ from the robot’s egocentric camera, and human hands differ visually from robot hands. Rather than masking hands or translating images across domains, HAT relies on diverse data and standard image augmentations such as color jittering and Gaussian blur. This keeps the training pipeline simple while allowing the frozen visual backbone to absorb moderate appearance shifts.

2.5 Experiments

Experiments use two humanoid robots equipped with 6-DOF Inspire dexterous hands. Humanoid A is a Unitree H1 and provides most of the robot demonstrations. Humanoid B is a Unitree H1-2 with different arm motor configurations and is used to evaluate few-shot transfer across humanoid embodiments. Both robots use active head motion and egocentric visual observations rather than wrist cameras.

Figure [Figure 2.3](#) should accompany the experimental setup because the embodiment differences between the two humanoids are central to the cross-embodiment evaluation. The baseline is ACT, an action-chunking transformer trained only on robot data with robot joint positions as state. HAT uses the same action-chunking architecture, but replaces the state encoder with the unified human-humanoid representation. Unless otherwise specified, HAT is co-trained on human and robot demonstrations. Each task policy is trained with approximately 250 to 400 robot demonstrations.

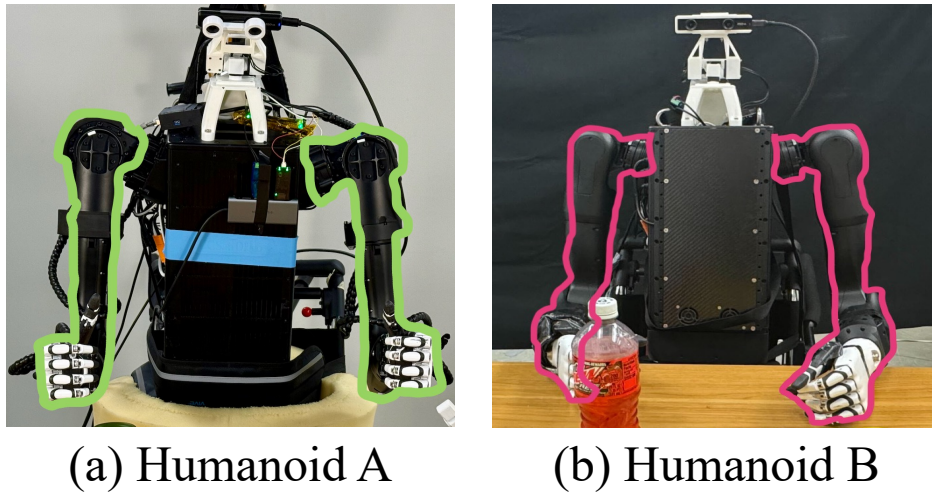


Figure 2.3. Humanoid platforms used in the HAT experiments. Humanoid A is the primary data collection platform, while Humanoid B has a different arm configuration and is used to evaluate few-shot cross-humanoid transfer.

Table 2.3. Autonomous execution success on Humanoid A. Co-training with human data preserves in-distribution performance and substantially improves out-of-distribution generalization across background, texture, object placement, and pouring settings.

Method	Human Data	Diff. Norm	Passing		Horizontal Grasp		Vertical Grasp		Pouring		Overall	
			I.D.	O.O.D.	I.D.	O.O.D.	I.D.	O.O.D.	I.D.	O.O.D.	I.D.	O.O.D.
ACT	No	N/A	19/20	36/60	8/10	7/30	7/20	15/70	8/10	1/10	42/60	59/170
HAT	Yes	No	17/20	51/60	9/10	11/30	14/20	30/70	5/10	5/10	45/60	97/170
HAT	Yes	Yes	20/20	52/60	8/10	12/30	13/20	29/70	8/10	8/10	49/60	101/170

The evaluation includes four manipulation settings: cup passing, horizontal grasping, vertical grasping, and pouring. In-distribution evaluation matches the robot demonstration setup. Out-of-distribution evaluation changes one factor at a time, using variations present in human data but absent from robot data. Cup passing tests background generalization, horizontal grasping tests object appearance and texture, vertical grasping tests object placement, and pouring tests changes in relative robot-table configuration.

Table [Table 2.3](#) shows the main result. Human data has a minor effect on in-distribution performance because the robot-only baseline already sees similar backgrounds, objects, and placements during teleoperation. The difference appears under distribution shift. HAT improves O.O.D. success from 59/170 to 101/170 overall when trained with human data and embodiment-specific normalization. This corresponds to large gains across background, object placement, texture, and pouring variations. The few-shot humanoid-transfer results further show that, when the embodiment changes and robot data are scarce, human data

Table 2.4. Background generalization in the cup passing task. The robot-only baseline is strongest on the in-distribution paper background, while HAT improves robustness across held-out backgrounds.

Method	Paper	Wooden	Red	Green	Overall
ACT	19/20	14/20	12/20	10/20	55/80
HAT	20/20	16/20	18/20	18/20	72/80

Table 2.5. Object appearance generalization in horizontal grasping. Box1 is observed in human demonstrations but not robot demonstrations, while Box2 and Can are harder out-of-distribution objects.

Method	Bottle	Box1	Box2	Can	Overall
ACT	8/10	5/10	1/10	1/10	16/40
HAT	8/10	7/10	1/10	4/10	21/40

acts as a larger cross-embodiment substrate rather than as a substitute for robot collection. The same theme appears in the few-shot humanoid transfer setting. Humanoid B has a different arm configuration and is collected in a separate environment, so the evaluation probes both embodiment shift and scene shift. With only a small number of Humanoid B demonstrations, co-training with human data and Humanoid A data improves success substantially over training on Humanoid B alone. The important point is that human data is not used as a surrogate for robot demonstrations; it is used as a distinct but aligned embodiment that expands the policy’s support.

2.5.1 Generalization Analysis

The gains in Table [Table 2.3](#) come from three main forms of generalization. In cup passing, robot demonstrations are collected on a limited set of table backgrounds, while human demonstrations include more background variation. Co-training improves robustness to tablecloth and scene appearance changes. In horizontal grasping, human data exposes the policy to more object appearances, improving performance on harder objects not covered by the robot demonstrations. In vertical grasping, human demonstrations cover a wider range of object positions, improving spatial generalization beyond the cells used for robot data collection.

Table [Table 2.4](#) isolates background variation in cup passing. HAT improves overall success from 55/80 to 72/80 and is especially stronger on the red and green backgrounds, which are out of distribution for robot-only training.

Table [Table 2.5](#) shows that human co-training improves grasping on several object appearances without reducing performance on the in-distribution bottle. The effect is strongest when the human dataset contains relevant object categories or visual priors for graspable regions.

2.5.2 Few-Shot Transfer to a New Humanoid

HAT also improves adaptation to Humanoid B, a robot with different arm configuration and a separate data collection environment. Figure [Figure 2.4](#) evaluates the three training regimes using 20 Humanoid B demonstrations. Humanoid B-only training achieves 26% success on the in-distribution object and near-zero success on all three held-out objects. Adding Humanoid A data improves the in-distribution and first held-out object to 73% and 52%, while adding PH²D further raises them to 86% and 70% and yields 55% and 10% success on the remaining held-out objects.

Figure [Figure 2.5](#) shows that this benefit persists as the target-robot dataset grows. At 20 Humanoid B demonstrations, co-training reaches 78% success versus 34% for Humanoid B-only training; at 50 demonstrations, the corresponding rates are 95% and 63%. The largest relative advantage occurs in the low-data regime, where five Humanoid B demonstrations yield 25% success with co-training and 1% without it.

2.5.3 Ablations

Two ablations identify which design choices make human-humanoid co-training effective. First, the policy must use a shared state-action representation. If the model receives robot joint positions for humanoid data and end-effector representations for human data, it can distinguish embodiments through the input format and overfit to the source domain. Second, human actions must be temporally aligned to robot execution speed. Without slowing human trajectories, the policy produces unstable rollouts because predicted action chunks mix fast human-like motion with slower teleoperation-like motion.

The results also show that human data improves robustness to background variation, object appearance, and object placement more strongly than it changes in-distribution behavior. That pattern is important: the value of PH²D is not simply additional demonstrations, but demonstrations whose variation is different from what robot teleoperation usually covers.

Table [Table 2.6](#) summarizes this result. The full design is not only a matter of adding more data; it requires aligning the representation and temporal structure of human and robot behavior. Figures [Figure 2.4](#) and [Figure 2.5](#) make the same point at the embodiment level: the human dataset helps most when the robot data are sparse and the target humanoid

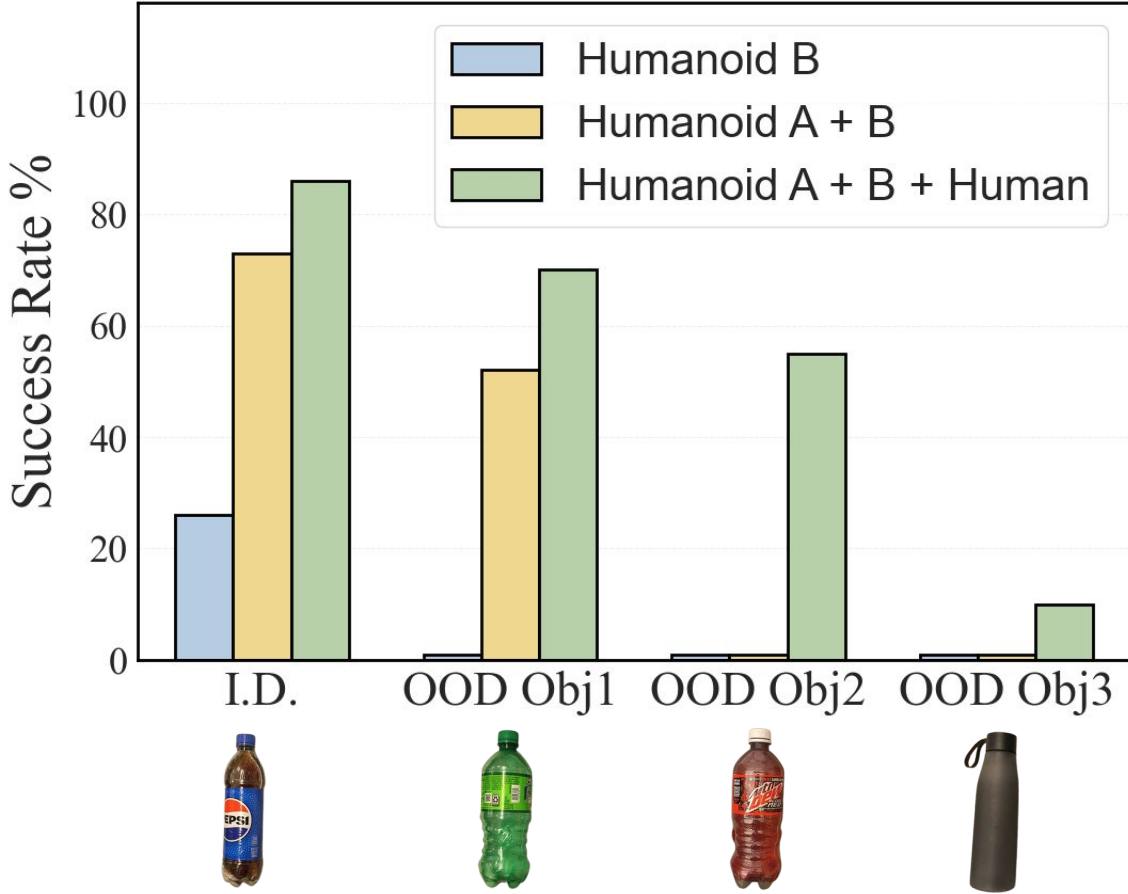


Figure 2.4. Few-shot cross-humanoid transfer across objects using 20 Humanoid B demonstrations. The in-distribution object is observed for Humanoid B; the held-out objects increase in difficulty and coverage. Co-training with Humanoid A data improves transfer, and adding PH²D provides the strongest performance across all object settings.

differs in arm configuration.

2.6 Conclusion

HAT demonstrates that task-oriented egocentric human demonstrations can serve as scalable cross-embodiment data for humanoid manipulation. PH²D provides human demonstrations with visual observations, head pose, wrist pose, and hand keypoints aligned to robot manipulation tasks. HAT then trains a transformer policy in a unified human-humanoid state-action space and retargets its predictions to robot actions.

The main result is that human co-training improves out-of-distribution robustness without sacrificing in-distribution performance. The improvement comes from variation that is easier to collect in human demonstrations than in robot teleoperation: backgrounds, object

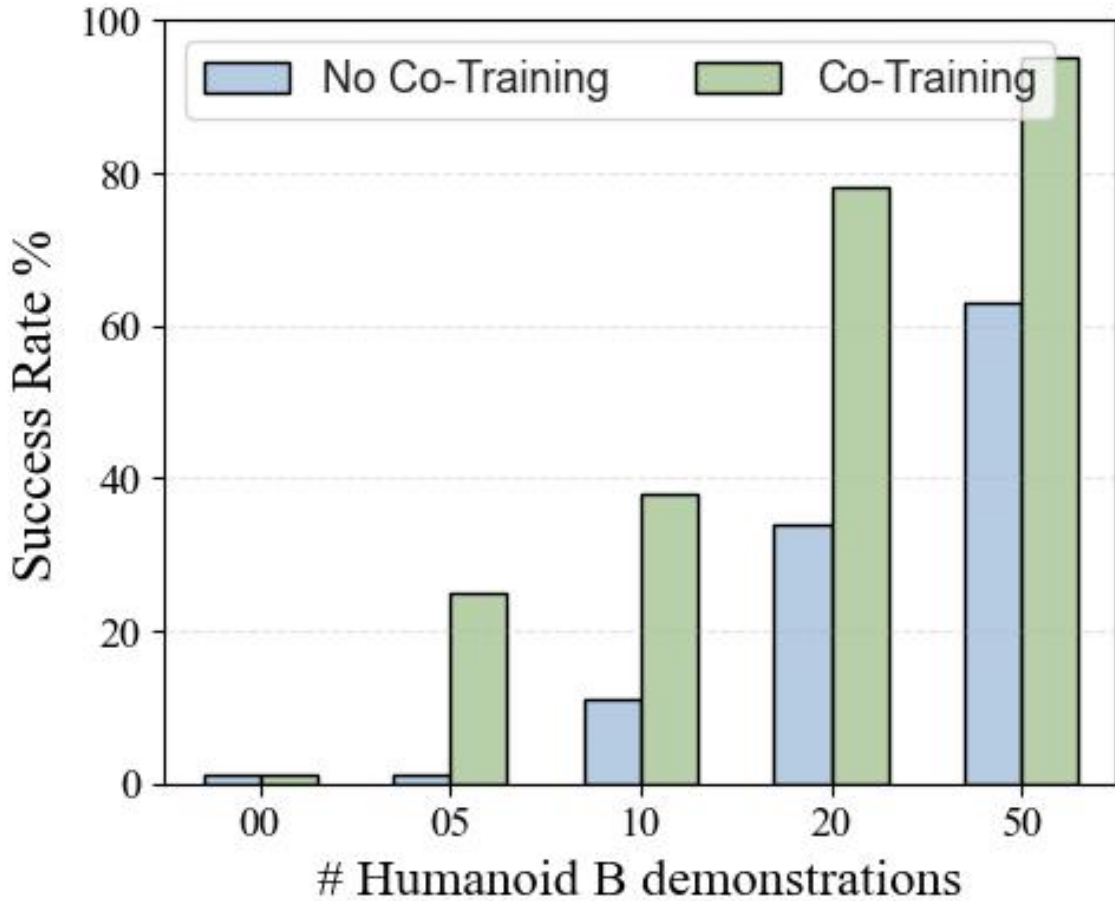


Figure 2.5. Few-shot cross-humanoid transfer as the number of Humanoid B demonstrations increases. Co-training holds the Humanoid A and PH²D datasets fixed and consistently outperforms Humanoid B-only training, especially with 5–20 target-robot demonstrations.

appearances, object placements, and task setups. At the thesis level, this chapter establishes that aligned human demonstrations can act as a direct supervision source for policy learning when the embodiment gap is moderate enough to be absorbed by a shared representation and retargeting layer.

This result also exposes the next bottleneck. HAT succeeds when human behavior can be summarized in a shared wrist-and-fingertip representation and the remaining mismatch can be handled by retargeting at execution time. Dexterous hand-object interaction is less forgiving: small errors in wrist reachability, finger contact, or object articulation can break the task even when the motion is semantically correct in human space. The next chapter therefore studies the complementary regime in which scalable human data remains useful, but direct co-training is no longer sufficient and embodiment constraints must instead be

Table 2.6. Ablation on unified state space and action speed alignment for vertical grasping in an out-of-distribution cell. Both ingredients are needed for stable transfer from human demonstrations.

Unified State Space	Slowed Human Actions	Success
Yes	No	1/10
No	Yes	0/10
Yes	Yes	4/10

introduced during motion generation before control.

Chapter 3

Embodiment-Aware Generation of Hand–Object Interactions

This chapter studies the thesis problem from a complementary angle to Chapter 2. Instead of asking whether aligned human demonstrations can directly supervise a humanoid policy, it asks how scalable human motion priors can remain useful when the embodiment gap is too large to defer entirely to post-hoc retargeting. Dexterous humanoid manipulation requires reference motions that are both semantically rich and feasible for the robot that must execute them. Human hand-object interaction data provides the right semantic structure: hands approach objects, establish contacts, articulate mechanisms, and coordinate bimanual motion in ways that match human environments. The difficulty is that these demonstrations are usually represented in human hand space, most commonly MANO [63], while humanoid robots differ in arm length, joint limits, wrist reachability, and finger morphology.

This chapter presents EMBODYHOI, an embodiment-aware pipeline for generating and executing dexterous hand-object interactions on humanoid robots. Instead of generating a human-space interaction and only later testing whether it can be retargeted, EMBODYHOI folds robot kinematics into the generative process. A differentiable inverse-kinematics (IK) module guides a text-conditioned hand-object interaction generator toward wrist trajectories that are reachable for a target humanoid, while also optimizing the robot base placement needed to execute the interaction. The resulting motions are retargeted to the robot and used as references for reinforcement learning with decoupled arm and hand rewards.

3.1 Introduction

Humanoid robots have made rapid progress in locomotion and whole-body control, but dexterous manipulation remains a more structured problem. A useful policy must coordinate whole-arm reaching, multi-finger contact, object articulation, and recovery under contact-

rich dynamics. Small mismatches in wrist pose or fingertip placement can prevent a hand from opening a box, grasping a handle, or maintaining contact with an articulated object. Within the thesis-level progression, this chapter addresses a harder setting than Chapter 2: direct human supervision is no longer enough, because dexterous hand-object interaction exposes a finer embodiment gap at the level of wrist reachability, finger contact, and object articulation. Generative motion models offer an attractive source of scalable manipulation references. A text-conditioned hand-object interaction generator can synthesize diverse motions from language and object geometry, extending beyond the demonstrations available for any single robot. However, most such models generate in human space. They produce MANO wrist and finger trajectories, then leave the embodiment gap to a post-hoc retargeting stage. This separation creates a brittle interface: a motion can be natural for a human hand yet place a humanoid arm near a kinematic singularity, outside its joint limits, or in a base pose that makes contact impossible.

Robot-space motion priors reduce this failure mode by training directly on retargeted trajectories for a fixed embodiment, as in DreamControl and DreamControl-v2 [64, 65]. This improves reference quality for a specific robot, but limits reuse across humanoid morphologies. Each new hand or arm requires new retargeted data and often a new generator. For dexterous manipulation, this cost is especially limiting because the relevant motion includes both coarse wrist placement and fine-grained finger-object contact.

EMBODYHOI takes a different position. It keeps a reusable human-space HOI prior and applies embodiment-specific guidance only at inference time. The method builds on MAD-HOI [66], a masked autoregressive diffusion model that generates hand-object interactions from text and articulated object geometry. During sampling, a differentiable IK objective scores decoded wrist trajectories against the target humanoid kinematics. The gradient of this objective steers the generator toward motions that preserve the semantic interaction while improving retargetability.

The full humanoid setting introduces a second structural issue: base placement. A floating human hand representation does not specify where the robot should stand relative to the object. A full-size humanoid and a smaller humanoid may need different base offsets to execute the same interaction without overextension. EMBODYHOI optimizes this base placement inside the IK objective, coupling wrist feasibility with the robot’s global pose.

This chapter therefore addresses the regime left open by Chapter 2: cases where scalable human data is available, but direct policy learning is insufficient because execution depends on finer embodiment details. The chapter develops three components around that goal. First, it introduces IK-guided sampling for text-conditioned hand-object motion generation. Second, it optimizes humanoid base placement jointly with arm IK, improving reachability

across robot morphologies. Third, it trains dexterous humanoid policies from the guided references using PPO [67], with separate reward and termination streams for the upper body and dexterous hands.

3.2 Related Work

Hand-object motion generation. Human hand-object interaction datasets such as ARCTIC [68], GRAB [69], HOI4D [70], OakInk [71], and DexYCB [72] capture bimanual dexterous behaviors with articulated objects, making them a natural source of manipulation priors. Text-conditioned HOI generators and motion priors use such data to synthesize motion from language or other conditioning signals [73–91], but their outputs are typically expressed in human body or hand representations. This makes them semantically rich, but not yet robot compatible.

Retargeting across embodiments. Retargeting maps human demonstrations to robot kinematics by matching end-effector poses, joint angles, or fingertip positions. Tools such as Pyroki [92] solve robot kinematic optimization for arm and wrist targets, while dexterous hand retargeting libraries map MANO hand motion to robot fingers [70, 93–96]. Functional retargeting methods such as DexMachina [97] further use contacts and task rewards to train policies from imperfect demonstrations, alongside related humanoid and teleoperation systems [68, 87, 92, 98–121]. These approaches are effective, but they operate after the reference motion is fixed. EMBODYHOI instead exposes robot reachability during generation, reducing the burden on downstream retargeting.

Generative priors for policy learning. Recent humanoid policy pipelines use generative motion priors to bootstrap physically grounded control. DreamControl [64, 65] uses generated human-like whole-body trajectories as references for reinforcement learning, and related work studies diffusion or motion-prior conditioning for humanoids and manipulation [4, 5, 122, 123]. EMBODYHOI shares the generate-then-track view, but targets dexterous hand-object manipulation and preserves cross-embodiment reuse by applying robot-specific IK guidance at inference time rather than retraining the generator.

3.3 Method

The method is organized around a thesis-level requirement complementary to HAT: if the data source remains human-space, embodiment constraints must be enforced before downstream control rather than deferred entirely to retargeting.

EMBODYHOI has three stages, shown in Figure [Figure 3.1](#). First, MAD-HOI generates a hand-object interaction from a language prompt and articulated object mesh. Second, an

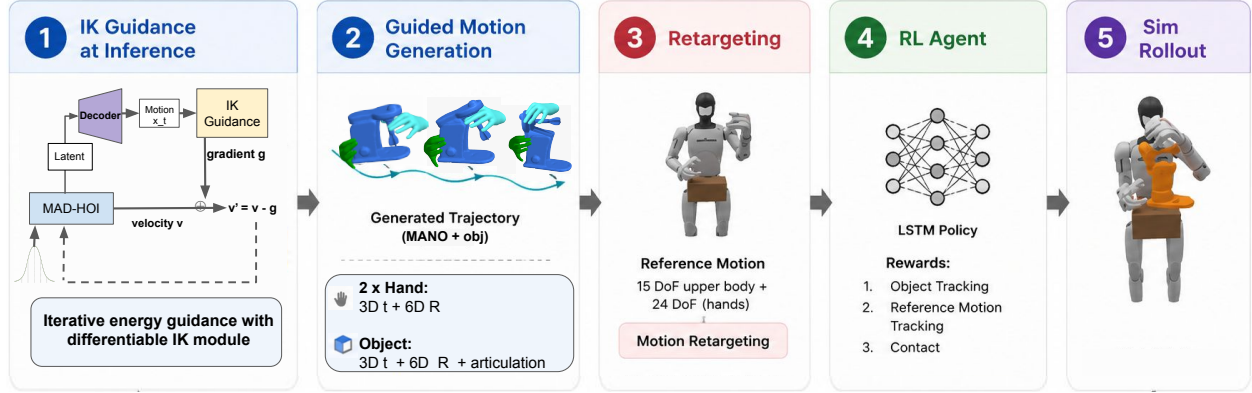


Figure 3.1. EMBODYHOI pipeline. A text-conditioned HOI generator proposes MANO hand and object trajectories. During sampling, differentiable IK guidance steers wrist motion toward the target humanoid kinematics and optimizes the robot base pose. The guided motion is retargeted and used to train a dexterous manipulation policy in simulation.

embodiment-aware IK objective guides the sampler toward wrist trajectories that the target humanoid can reach. Third, the guided interaction is retargeted to robot joint space and used to train a PPO policy for dexterous manipulation.

Text-conditioned HOI generation. MAD-HOI generates an object trajectory and per-hand MANO motion. The object state at time t contains translation, rotation, and articulation, written as $(T_{\text{obj}}^t, R_{\text{obj}}^t, \alpha_{\text{obj}}^t)$. Each hand state contains wrist translation and articulated finger pose, written as (τ_h^t, ϕ_h^t) . The model represents object, left-hand, and right-hand motion as decoupled latent tokens, which allows it to generate long, compositional hand-object interactions while preserving detailed hand motion.

The generator is trained in human hand space, so its samples are not constrained by robot arm kinematics. A generated wrist path may require the robot to overextend, pass through a singular configuration, or stand in an infeasible location. EMBODYHOI addresses this at sampling time rather than after generation.

Compared to earlier hand-object generators that focus on short atomic clips or tightly coupled object-hand states, this formulation is more suitable for robotics because it can represent longer manipulation episodes and preserve the object motion needed for downstream retargeting. The ARCTIC benchmark provides the interaction diversity needed for that setting, with 4.6k atomic motion sequences spanning 11 articulated objects.

IK-guided sampling. Let $\mathbf{x}$ denote the latent state being denoised by the flow-matching sampler, and let $\mathbf{z}$ denote the text and context conditioning. Standard sampling follows the learned velocity field $\mathbf{v}_\theta(\mathbf{x}, t, \mathbf{z})$. EMBODYHOI adds an embodiment-specific energy term:

$$\dot{\mathbf{x}} = \mathbf{v}_\theta(\mathbf{x}, t, \mathbf{z}) - w \nabla_{\mathbf{x}} E(\mathbf{x}),$$

where E is the IK residual and w controls guidance strength.

At each denoising step, the partially decoded hand motion provides target wrist poses. Guidance is applied only to the hand latent tokens, leaving the decoupled object tokens available to preserve the generated interaction. For each hand h , the robot arm Jacobian

$$\mathbf{J}_h(\mathbf{q}^t) = \frac{\partial \text{vec}(\mathbf{T}_{ee,h}^t)}{\partial \mathbf{q}^t}$$

measures local kinematic conditioning. Low minimum singular values and high condition numbers indicate proximity to singularities. Guided sampling uses 20 fourth-order Runge–Kutta integration steps; each per-timestep IK solve uses up to 10 Levenberg–Marquardt iterations. Its primary objective penalizes wrist-pose error, joint-limit violations, rest-pose deviation, temporal discontinuities, object-clearance violations, and excessive base displacement. A secondary null-space update then favors higher arm manipulability without changing the wrist target.

The result is an energy-guided sampler that preserves the learned HOI prior while biasing only the parts of the latent sequence that influence hand motion. In practice, this matters because the object tokens should remain available for compositional generation while the hand tokens absorb embodiment-specific corrections.

Base placement optimization. For humanoids, reachability depends not only on arm joints but also on where the robot stands. EMBODYHOI jointly optimizes arm joints and base translation at each timestep, with the base orientation fixed and translation regularized toward a shared prior. The optimized per-frame base estimates are then averaged to produce the fixed execution pose, which was more stable than a trajectory-wide shared-base solve. This lets the same human-space interaction adapt to robots with different scales and arm lengths.

Retargeting and policy learning. Unless otherwise specified, the target platform is a Unitree G1 humanoid equipped with XHands. The robot controls 15 upper-body degrees of freedom for torso and arms, with a fixed lower body, and 12 degrees of freedom per hand. Guided wrist trajectories are retargeted with Pyroki [92]; MANO finger articulation is mapped to the dexterous hands using dexterous hand retargeting [93]. Contact targets are retargeted following DexMachina [97].

The downstream policy is an LSTM trained with PPO in simulation. It observes robot state, object state, and sparse future references, then outputs a 39-dimensional hybrid action: delta targets for 15 torso and arm degrees of freedom and absolute targets for 24 finger degrees

of freedom. The reward is decoupled into an upper-body imitation stream over 15 degrees of freedom and a dexterous hand stream over 24 finger degrees of freedom. This separation is important because arm tracking is dominated by coarse reachability and stability, while finger control requires precise contact alignment. Termination is also decoupled: upper-body failure is measured by maximum per-joint deviation, while hand failure is measured by mean finger deviation. Once terminated episodes reach 80% of the task horizon on average, training halves the entropy coefficient and disables random-state initialization to emphasize reliable execution.

This design matches the control structure of the task. Upper-body control is mostly about getting the wrist and torso to the right workspace region, whereas hand control is about preserving contact and object-specific articulation once the arm is in place. The policy is memory-based, trained with PPO using rl-games [124] in Genesis [125] with heavily parallel simulation, and conditioned on sparse future references rather than only current state. Appendix B contains the full guidance, observation, reward, termination, and hyperparameter specifications.

3.4 Experimental Results

The results address whether IK-guided generation improves kinematic conditioning, preserves the quality of the HOI prior, and supports dexterous policy learning.

Kinematic conditioning. The main purpose of IK guidance is to steer generated wrist trajectories away from singular configurations before retargeting. Figure Figure 3.2 compares unguided and IK-guided generations for the Unitree G1. For each trajectory, the minimum singular value and maximum condition number of the upper-body kinematic chain are computed over time. Guided generations achieve higher minimum singular values and lower condition numbers, indicating better-conditioned arm configurations.

This metric is especially relevant when the robot and human morphologies diverge. Smaller differences in link proportions may be tolerable after retargeting, but larger mismatches tend to amplify singular behavior. The strongest gains appear where the embodiment gap is larger.

Generation quality. Guidance should not destroy the learned hand-object interaction prior. Table Table 3.1 compares unguided MAD-HOI samples with IK-guided samples on 419 ARCTIC examples. IK guidance introduces a modest distributional tradeoff: FID and KID increase, and diversity decreases slightly. However, text-motion retrieval accuracy remains nearly unchanged, contact is comparable, and penetration improves slightly. The guided samples therefore become more robot-compatible without collapsing semantic align-

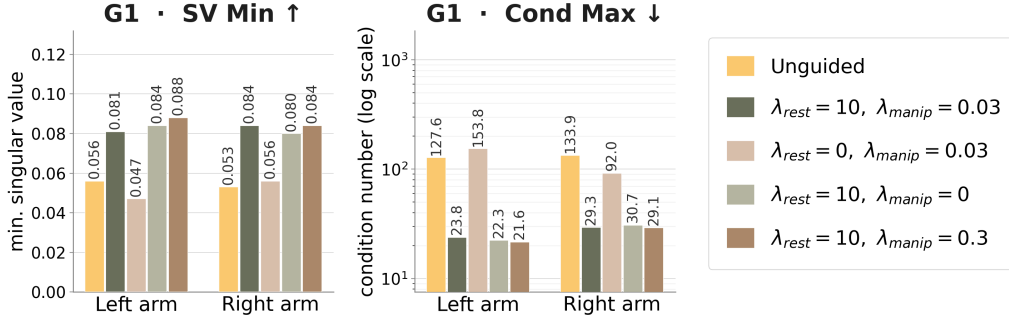


Figure 3.2. Kinematic conditioning of unguided and IK-guided generations for the Unitree G1. IK guidance increases the minimum Jacobian singular value and decreases the maximum condition number, reducing proximity to singular configurations.

Table 3.1. Generation quality for unguided and IK-guided samples on 419 ARCTIC examples.

Method	FID ↓	KID ↓	Diversity ↑	Top-3 Acc. ↑	Match ↓	Contact ↑	Pen. < 1cm ↓	Jitter ↓
Unguided	0.474	0.005	20.315	0.871	3.487	83.521	12.224	0.287
IK-Guided	0.615	0.009	19.380	0.869	3.640	83.147	11.512	0.293

ment or hand-object structure.

That tradeoff is expected rather than problematic. The guidance objective intentionally perturbs the original human-space sample distribution, so some degradation in generative metrics is the cost of moving trajectories toward regions that a humanoid can execute. The important pattern is that the metrics tied most directly to semantic structure, such as retrieval and contact, remain nearly unchanged. Guidance is therefore acting selectively: it shifts the wrists toward more feasible configurations without rewriting the underlying interaction.

IK-guidance ablation. Table [Table 3.2](#) isolates the rest-pose and null-space manipulability terms. The rest-pose prior regularizes the motion toward smoother, less distorted samples, while the manipulability term imposes a soft embodiment-specific posture preference. Their default weights are $(\lambda_{rest}, \lambda_{manip}) = (10, 0.03)$.

The ablation quantifies the expected cost of embodiment-aware regularization. Removing the rest-pose term worsens distributional and semantic metrics, while increasing the manipulability weight by an order of magnitude sharply reduces diversity and worsens FID and KID. The zero-manipulability setting has stronger generic generation scores, so the default retains a small nonzero weight as a soft kinematic preference rather than maximizing it.

End-to-end executability. The most direct test is downstream manipulation success. Table [Table 3.3](#) compares policies trained from guided and unguided samples on five ARCTIC

Table 3.2. Compact view of the IK-guided sampling ablation. The default setting balances rest-pose and manipulation terms; stronger manipulation guidance can increase retrieval while changing the distribution more sharply.

$(\lambda_{\text{rest}}, \lambda_{\text{manip}})$	FID ↓	KID ↓	Diversity ↑	Top-3 Acc. ↑	Match ↓
(10, 0.03)	0.615	0.009	19.380	0.869	3.640
(0, 0.03)	0.690	0.013	19.900	0.858	3.760
(10, 0)	0.545	0.006	19.850	0.872	3.570
(10, 0.30)	0.820	0.022	17.900	0.835	3.920

tasks whose unguided references have poor kinematic conditioning. All end-to-end results are simulated Unitree G1 rollouts. Guided samples improve success on every task, with the average success rate rising from 47% to 76%. This is a targeted stress test of whether generation-time guidance improves the usefulness of the reference, not the policy optimizer. For each evaluation trajectory, success is averaged over 20 episodes with independently sampled random seeds; an episode must remain within 0.05 cm position, 20° orientation, and 10° articulation error from the reference at every timestep. The source reports aggregate success rates only, so seed-level variance is not available.

This result is important because PPO is not creating dexterous behavior from scratch. The policy succeeds when the reference trajectory already lies inside a reachable, well-conditioned part of the robot workspace. Guided samples reduce upstream geometric errors, so reinforcement learning spends less capacity repairing impossible or near-singular motions and more capacity refining contact execution. The uneven gains across tasks support that interpretation. The small improvement on closing the microwave suggests that this task was already relatively reachable in the unguided setting, leaving less room for kinematic correction. By contrast, the much larger gains on bimanual box and mixer tasks indicate that embodiment mismatch is a stronger bottleneck when both reachability margins and cross-hand coordination are tight.

Reward design. Figure [Figure 3.3](#) ablates the reward structure used for downstream policy learning. Imitation-only policies track references but often fail to complete contact-sensitive tasks. Task-only rewards provide weak exploration signals in the high-dimensional hand-action space. A shared imitation reward over arms and fingers creates an imbalance between coarse reaching and fine dexterous contact. The decoupled reward used by EMBODY-HOI yields the highest success because it lets the policy learn stable upper-body tracking and fine finger control on separate scales.

Table 3.3. Mean simulated Unitree G1 reference-tracking success of policies trained from unguided and IK-guided generated samples. Each evaluation trajectory is run for 20 independently seeded episodes and then averaged across the task’s trajectories. Success requires every timestep to remain within 0.05 cm position, 20° orientation, and 10° articulation error of the reference. Seed-level variance was not reported in the source experiments.

Task	Unguided Success (%)	IK-Guided Success (%)
Close box with both hands	31	77
Close microwave with left hand	65	68
Lift box with both hands	56	89
Open mixer with both hands and close mixer with both hands	21	63
Lift mixer with both hands and place mixer with both hands	62	85
Average	47	76

The reward design table separates upper-body tracking, hand tracking, contact terms, and termination conditions for exactly this reason. Without that decomposition, the policy tends to either overfit to coarse wrist motion or get stuck in unstable finger exploration.

Qualitative cross-embodiment adaptation. Because guidance is applied at inference time, the same human-space generator can be adapted to different humanoids. Figure [Figure 3.4](#) visualizes this process for Unitree H1-2 and Unitree G1 [126, 127]. H1-2 is a full-size humanoid with longer arms, while G1 is smaller with shorter reach. Starting from the same interaction, EMBODYHOI selects different base placements for each robot, moving the embodiment into a more reachable configuration before retargeting. The H1-2 singularity analysis in Appendix B shows less pronounced changes than on G1, indicating that the kinematic effect depends on the target embodiment and task distribution.

This comparison is informative because the two robots differ not just in scale but also in arm proportions and mechanical reach. It visualizes embodiment adaptation at inference time rather than a robot-specific generator. However, it is not a controlled cross-embodiment policy evaluation: the reported execution metrics in Table [Table 3.3](#) are G1-only, and H1-2 is assessed only through qualitative adaptation and kinematic conditioning.

3.5 Conclusion

EMBODYHOI shows that embodiment constraints can be introduced during hand-object motion generation without retraining a separate generator for each robot. By differentiating

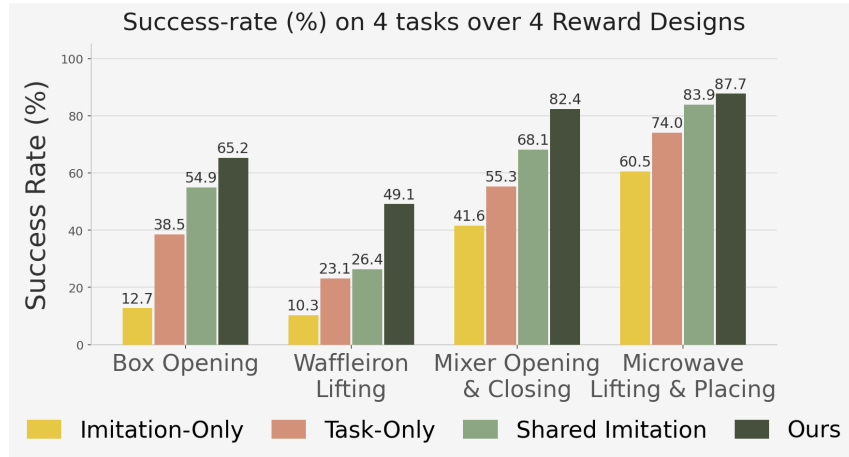


Figure 3.3. Reward design ablation for downstream policy learning. Decoupling arm and hand imitation rewards improves success relative to imitation-only, task-only, and shared-imitation variants.

through an IK objective at inference time, the method steers a reusable human-space HOI prior toward wrist trajectories that are easier to retarget to humanoid embodiments. Joint base placement optimization further adapts the same interaction to robots with different scales and arm lengths.

The resulting references preserve most of the semantic and contact quality of the underlying generator while improving kinematic conditioning. When retargeted and used for policy learning, they support dexterous manipulation of articulated objects with decoupled arm and hand rewards. At the thesis level, this chapter complements Chapter 2 by showing that scalability cannot always come from direct co-training alone. When the embodiment gap becomes too fine-grained for a shared policy representation, scalable human data must instead be filtered through embodiment-aware generation before control. The remaining gap is dynamic: IK guidance reasons about reachability and singularities, while contact stability and actuation limits are resolved downstream by simulation-based reinforcement learning. In this sense, the chapter supports the complementary half of the thesis claim introduced in Chapter 1: when direct alignment of human demonstrations is no longer sufficient, human motion priors remain useful only if embodiment constraints are introduced before execution. Together, Chapters 2 and 3 show that scalable human data becomes valuable for robot manipulation only after alignment to the target robot body, whether that alignment occurs in the policy representation or in the motion-generation process.

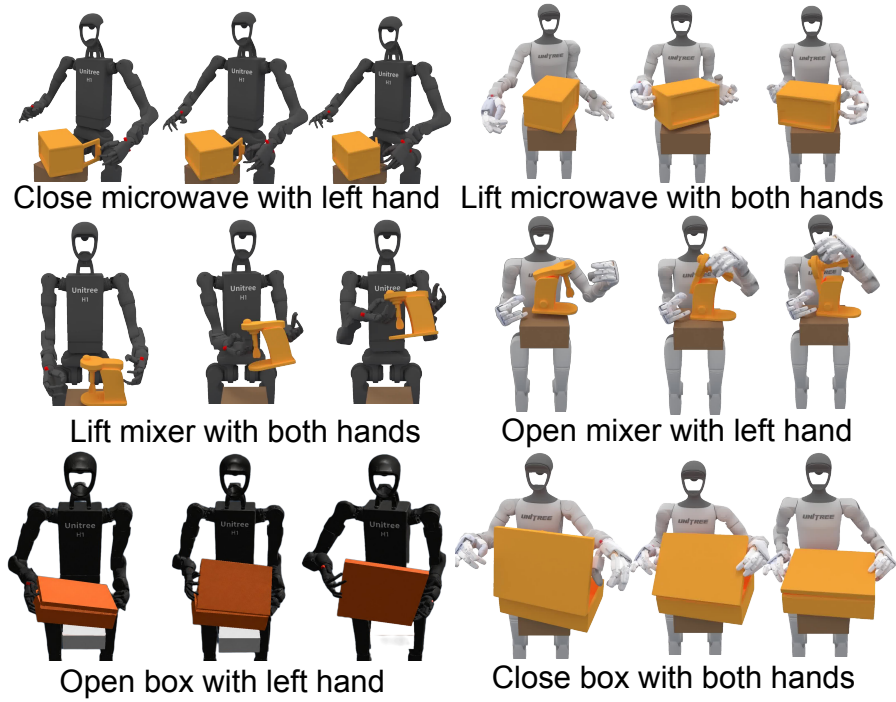


Figure 3.4. Qualitative cross-embodiment adaptation for Unitree H1-2 and Unitree G1. The same generated interaction is guided toward different base placements and wrist trajectories according to each humanoid’s morphology. This visualization is not an H1-2 policy-execution evaluation.

Chapter 4

Conclusion and Future Work

This thesis studied a central problem in robot learning: how to scale manipulation beyond robot-only data collection without losing the embodiment-specific structure required for execution. For humanoid and dexterous systems, this problem is especially acute. The most useful supervision must capture contact, reachability, wrist placement, finger coordination, and object motion, yet collecting such data directly on robots is slow, expensive, and tightly coupled to a specific hardware platform. The central claim of the thesis is that human behavior is the right scalable source of supervision, provided that it can be converted into forms that remain faithful to the target robot body.

Chapters 2 and 3 developed two complementary solutions to this bottleneck. Chapter 2 showed that task-oriented egocentric human demonstrations can serve as direct cross-embodiment supervision for humanoid manipulation. By placing humans and humanoids in a unified state-action representation, HAT enables co-training across human and robot demonstrations and improves robustness to variation in backgrounds, object appearances, and object placements. This result supports the view that aligned human demonstrations can expand the diversity of manipulation experience without requiring a robot for every demonstration.

Chapter 3 addressed a second setting in which scalable manipulation data comes not from direct demonstrations, but from generative hand-object interaction priors. EMBODYHOI showed that this source is useful only when embodiment is accounted for before downstream control. By guiding HOI generation with differentiable inverse kinematics and joint base-placement optimization, the method preserves the semantic richness of human-space motion priors while producing references that are more reachable and better conditioned for humanoid execution. This result supports the complementary claim that scalable manipulation requires not only more motion data, but motion data shaped by the constraints of the executing body.

Taken together, these chapters suggest a broader lesson for robot learning. The main obstacle is not simply the quantity of data, but the mismatch between abundant human behavior

and executable robot behavior. HAT and EMBODYHOI resolve this mismatch in different ways. HAT aligns demonstrations in a shared representation for direct policy learning, while EMBODYHOI aligns generated interactions by injecting embodiment constraints during sampling. Across both projects, the common principle is that scale becomes useful only after alignment.

4.1 Future Work

The most natural next steps are the ones that target the remaining alignment gaps exposed by the two systems.

From task-oriented collection to web-scale human video. HAT shows the value of aligned human demonstrations, but PH²D still relies on intentional task-oriented collection with egocentric cameras and hand supervision. A natural next step is to use much broader human video sources during a pretraining stage, but to focus that source specifically on manipulation rather than training on arbitrary internet video. The goal is for the model to first learn generic structure about how humans approach objects, establish contact, regrasp, manipulate articulated parts, and coordinate bimanual interaction across diverse environments. The role of a web-scale data collection pipeline would then be to harvest manipulation-centric episodes from internet video and convert them into weak but scalable supervision through task segmentation, hand-motion estimation, viewpoint cues, and object state-change signals. That generic pretraining could later be post-trained or fine-tuned on aligned datasets such as PH²D, which provide the embodiment-aware structure needed for deployment. Relative to PH²D, the challenge is not just scale but structure: the web data are less aligned, less labeled, and visually less controlled, yet they cover a much broader range of objects and environments. This direction would test whether the thesis claim still holds at internet scale, namely that large human datasets become valuable for robot learning only after they are transformed into embodiment-relevant training signals.

Beyond kinematic alignment. EMBODYHOI improves executability by guiding generation with inverse kinematics and base placement, but the remaining failures are often dynamic and contact-specific rather than purely geometric. A generated interaction can be reachable and still be difficult to execute because of contact timing, object inertial effects, limited hand strength, or mismatch between human finger motion and the target dexterous hand. This limitation is especially visible in tasks that require fine-grained contact, where a motion that is semantically correct in MANO space may still fail after retargeting because the robot hand cannot reproduce the same contact pattern reliably. One practical direction is therefore a targeted hybrid regime: use human demonstrations and human-space genera-

tive priors for broad task coverage, then reserve a small amount of robot data for the phases where embodiment mismatch matters most. Rather than collecting robot demonstrations uniformly, a more efficient strategy would be to use generated references or human-trained policies to identify failure modes, such as unstable fingertip contact, poor wrist orientation near grasp, or articulation phases that repeatedly break after retargeting, and then collect robot data specifically for those segments. That hybrid approach would likely improve execution on contact-sensitive tasks, although it also weakens the original goal of relying on human-only data and reusable generative priors. A longer-term alternative is to introduce richer contact-aware signals into the guidance objective itself, for example by penalizing unstable contacts, large tracking torques, or reference motions that repeatedly fail in downstream PPO training. The broader goal is to make the generated references and retargeting process accurate enough that reinforcement learning is no longer needed as a corrective stage for residual embodiment and contact errors. In that regime, RL would no longer be responsible for repairing failures introduced upstream, and generation would move closer to producing directly executable robot behavior.

In summary, this thesis argues that the path to more general humanoid manipulation lies in scaling data without abandoning embodiment. Human behavior is abundant, diverse, and structurally close to the tasks robots must perform, but it becomes useful for robot learning only after careful alignment. By showing this through both cross-embodiment imitation and embodiment-aware motion generation, the thesis offers a concrete path toward more scalable manipulation systems.

Bibliography

- [1] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *CoRL*, 2023.
- [2] 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.
- [3] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *RSS*, 2024.
- [4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [5] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.
- [6] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. In *ECCV*, 2024.
- [7] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.
- [8] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *RSS*, 2023.
- [9] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *arXiv preprint arXiv:2212.04498*, 2023.

- [10] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5954–5961. IEEE, 2023.
- [11] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning (CoRL)*, 2024.
- [12] 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.
- [13] 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.
- [14] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *RSS*, 2022.
- [15] 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.
- [16] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.
- [17] Neha Das, Sarah Bechtel, Todor Davchev, Dinesh Jayaraman, Akshara Rai, and Franziska Meier. Model-based inverse reinforcement learning from visual demonstrations. In *Conference on Robot Learning*, pages 1930–1942. PMLR, 2021.
- [18] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.
- [19] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from” in-the-wild” human videos. *arXiv preprint arXiv:2103.16817*, 2021.
- [20] Jangwon Lee and Michael S Ryoo. Learning robot activities from first-person human videos using convolutional future regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–2, 2017.

- [21] Kyuhwa Lee, Yanyu Su, Tae-Kyun Kim, and Yiannis Demiris. A syntactic approach to robot imitation learning using probabilistic activity grammars. *Robotics and Autonomous Systems*, 61(12):1323–1334, 2013.
- [22] Anh Nguyen, Dimitrios Kanoulas, Luca Muratore, Darwin G Caldwell, and Nikos G Tsagarakis. Translating videos to commands for robotic manipulation with deep recurrent neural networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3782–3788. IEEE, 2018.
- [23] Jonas Rothfuss, Fabio Ferreira, Eren Erdal Aksoy, You Zhou, and Tamim Asfour. Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution. *IEEE Robotics and Automation Letters*, 3(4):4007–4014, 2018.
- [24] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Robot learning manipulation action plans by” watching” unconstrained videos from the world wide web. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.
- [25] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.
- [26] Sören Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, and Pierre Sermanet. Online object representations with contrastive learning. *arXiv preprint arXiv:1906.04312*, 2019.
- [27] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.
- [28] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.
- [30] 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.

- [31] 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. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [32] Ri-Zhao Qiu, Yuchen Song, Xuanbin Peng, Sai Aneesh Suryadevara, Ge Yang, Minghuan Liu, Mazeyu Ji, Chengzhe Jia, Ruihan Yang, Xueyan Zou, et al. Wildlma: Long horizon loco-manipulation in the wild. *arXiv preprint arXiv:2411.15131*, 2024.
- [33] Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. In *ICRA*, 2025.
- [34] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024.
- [35] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024.
- [36] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se (3) pose trajectory diffusion for object-centric manipulation. *arXiv preprint arXiv:2411.00965*, 2024.
- [37] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *CoRL*, 2024.
- [38] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *arXiv preprint arXiv:2409.20537*, 2024.
- [39] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. *arXiv preprint arXiv:2402.19249*, 2024.
- [40] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024.
- [41] Jonathan Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.

- [42] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.
- [43] Tim Franzmeyer, Philip Torr, and João F Henriques. Learn what matters: cross-domain imitation learning with task-relevant embeddings. *Advances in Neural Information Processing Systems*, 35:26283–26294, 2022.
- [44] Ali Ghadirzadeh, Xi Chen, Petra Poklukar, Chelsea Finn, Mårten Björkman, and Danica Kragic. Bayesian meta-learning for few-shot policy adaptation across robotic platforms. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1274–1280. IEEE, 2021.
- [45] Tanmay Shankar, Yixin Lin, Aravind Rajeswaran, Vikash Kumar, Stuart Anderson, and Jean Oh. Translating robot skills: Learning unsupervised skill correspondences across robots. In *International Conference on Machine Learning*, pages 19626–19644. PMLR, 2022.
- [46] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555. PMLR, 2023.
- [47] Zhao-Heng Yin, Lingfeng Sun, Hengbo Ma, Masayoshi Tomizuka, and Wu-Jun Li. Cross domain robot imitation with invariant representation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 455–461. IEEE, 2022.
- [48] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*, pages 537–546. PMLR, 2022.
- [49] Grace Zhang, Linghan Zhong, Youngwoon Lee, and Joseph J Lim. Policy transfer across visual and dynamics domain gaps via iterative grounding. *arXiv preprint arXiv:2107.00339*, 2021.
- [50] Qiang Zhang, Tete Xiao, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Learning cross-domain correspondence for control with dynamics cycle-consistency. *arXiv preprint arXiv:2012.09811*, 2020.
- [51] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via ego-centric video. *arXiv preprint arXiv:2410.24221*, 2024. URL <https://arxiv.org/abs/2410.24221>.

- [52] Li-Heng Lin, Yuchen Cui, Amber Xie, Tianyu Hua, and Dorsa Sadigh. Flowretrieval: Flow-guided data retrieval for few-shot imitation learning. *arXiv preprint arXiv:2408.16944*, 2024.
- [53] Juntao Ren, Priya Sundareshan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *arXiv preprint arXiv:2501.06994*, 2025.
- [54] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023.
- [55] Jonathan Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.
- [56] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. *arXiv preprint arXiv:2402.19249*, 2024.
- [57] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *CoRL*, 2024.
- [58] 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.
- [59] 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.
- [60] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via ego-centric video. *arXiv preprint arXiv:2410.24221*, 2024.
- [61] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *CoRL*, 2024.
- [62] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [63] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 36(6), 2017.

- [64] Dvij Kalaria, Sudarshan S Harithas, Pushkal Katara, Sangkyung Kwak, Sarthak Bhagat, Shankar Sastry, Srinath Sridhar, Sai Vemprala, Ashish Kapoor, and Jonathan Chung-Kuan Huang. Dreamcontrol: Human-inspired whole-body humanoid control for scene interaction via guided diffusion. *arXiv preprint arXiv:2509.14353*, 2025.
- [65] Sudarshan Harithas, Sangkyung Kwak, Pushkal Katara, Srujan Deolasee, Dvij Kalaria, Srinath Sridhar, Sai Vemprala, Ashish Kapoor, and Jonathan Chung-Kuan Huang. Dreamcontrol-v2: Simpler and scalable autonomous humanoid skills via trainable guided diffusion priors. *arXiv preprint arXiv:2604.00202*, 2026.
- [66] Anonymized author list. Mad-hoi: Masked autoregressive diffusion for generating hand object interactions from text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. In Submission.
- [67] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [68] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023.
- [69] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.
- [70] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [71] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [72] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

- [73] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [74] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.
- [75] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.
- [76] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [77] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.
- [78] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [79] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Mo-Mask: Generative masked modeling of 3D human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [80] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [81] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2023.
- [82] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision (ECCV)*, 2024.
- [83] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024.

- [84] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. HOI-Diff: Text-driven synthesis of 3D human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023.
- [85] Christian Diller and Angela Dai. CG-HOI: Contact-guided 3D human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [86] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. ContactGen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [87] Xueyi Liu and Li Yi. GeneOH diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.
- [88] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3D motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [89] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. DiffH2O: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia Conference Papers*, 2024.
- [90] Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. *Advances in Neural Information Processing Systems*, 38:166582–166612, 2026.
- [91] Zhenhao Zhang, Jiabin Liu, Ye Shi, and Jingya Wang. Unihm: Unified dexterous hand manipulation with vision language model. *arXiv preprint arXiv:2603.00732*, 2026.
- [92] Chung Min Kim, Brent Yi, Hongsuk Choi, Yi Ma, Ken Goldberg, and Angjoo Kanazawa. Pyroki: A modular toolkit for robot kinematic optimization. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1312–1319. IEEE, 2025.
- [93] 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. *arXiv preprint arXiv:2307.04577*, 2023.
- [94] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*, 2022.

- [95] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [96] RobotEra. XHAND1 dexterous hand, 2024. Product page. [Online]. Available: <https://www.robotera.com/en/goods1/4.html>.
- [97] Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. Dexmachina: Functional retargeting for bimanual dexterous manipulation. *arXiv preprint arXiv:2505.24853*, 2025.
- [98] 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.
- [99] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9989–9996. IEEE, 2025.
- [100] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.
- [101] Yunshen Wang, Shaohang Zhu, Peiyuan Zhi, Yuhan Li, Jiaxin Li, Yong-Lu Li, Yuchen Xiao, Xingxing Wang, Baoxiong Jia, and Siyuan Huang. Omnixtreme: Breaking the generality barrier in high-dynamic humanoid control. *arXiv preprint arXiv:2602.23843*, 2026.
- [102] Tairan He, Zi Wang, Haoru Xue, Qingwei Ben, Zhengyi Luo, Wenli Xiao, Ye Yuan, Xingye Da, Fernando Castañeda, Shankar Sastry, Changliu Liu, Guanya Shi, Linxi Fan, and Yuke Zhu. Viral: Visual sim-to-real at scale for humanoid loco-manipulation. *arXiv preprint arXiv:2511.15200*, 2025.
- [103] Jhen Hsieh, Kuan-Hsun Tu, Kuo-Han Hung, and Tsung-Wei Ke. Dexman: Learning bimanual dexterous manipulation from human and generated videos. *arXiv preprint arXiv:2510.08475*, 2025.
- [104] Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latentmoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17416–17425, 2025.

- [105] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6991–7003, 2025.
- [106] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems*, 37:2161–2184, 2024.
- [107] Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. Spider: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484*, 2025.
- [108] Sungjae Park, Homanga Bharadhwaj, and Shubham Tulsiani. Demodiffusion: One-shot human imitation using pre-trained diffusion policy. *arXiv preprint arXiv:2506.20668*, 2025.
- [109] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.
- [110] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.
- [111] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2878–2888, 2025.
- [112] Sharpa Robotics. Sharpa wave. <https://www.sharpa.com/pages/wave>, 2024. Accessed: 2026-01-31.
- [113] Kenneth Shaw and Deepak Pathak. Leap hand v2: Dexterous, low-cost anthropomorphic hybrid rigid soft hand for robot learning. In *2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS)*, 2024.
- [114] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

- [115] Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17540–17549, 2025.
- [116] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025.
- [117] Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*, 2025.
- [118] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.
- [119] Haoru Xue, Tairan He, Zi Wang, Qingwei Ben, Wenli Xiao, Zhengyi Luo, Xingye Da, Fernando Castañeda, Guanya Shi, Shankar Sastry, Linxi ”Jim” Fan, and Yuke Zhu. Opening the sim-to-real door for humanoid pixel-to-action policy transfer. *arXiv preprint arXiv:2512.01061*, 2025.
- [120] Zhecheng Yuan, Tianming Wei, Langzhe Gu, Pu Hua, Tianhai Liang, Yuanpei Chen, and Huazhe Xu. Hermes: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation. *arXiv preprint arXiv:2508.20085*, 2025.
- [121] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023.

- [122] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024.
- [123] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR, 2025.
- [124] Denys Makoviichuk and Viktor Makoviyshuk. rl-games: A high-performance framework for reinforcement learning. https://github.com/Denys88/rl_games, May 2021.
- [125] Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond. <https://github.com/Genesis-Embodied-AI/Genesis>, December 2024.
- [126] Unitree Robotics. H1 / H1-2 Universal Humanoid Robot. <https://unitree-robot.com/h1/>, 2026. Accessed: 2026-05-28.
- [127] Unitree Robotics. G1 Humanoid Robot. <https://www.unitree-robot.com/mobile/g1>, 2026. Accessed: 2026-05-28.

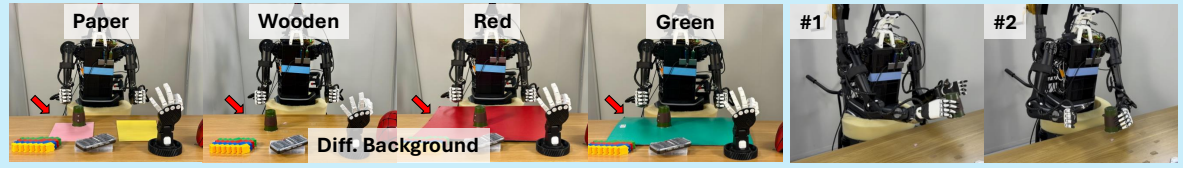


Figure A.1. The robot performs the **cup passing** task across four different backgrounds. The left side shows the four background variations, while the right side illustrates the two passing directions: (#1 - Right hand passes the cup to the left hand, #2 - Left hand passes the cup to the right hand).

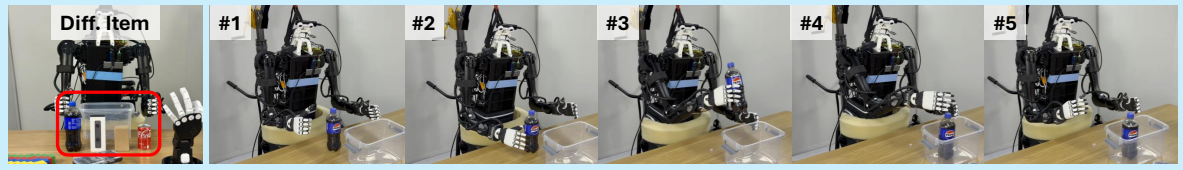


Figure A.2. The robot performs the **horizontal grasping** task with four different items: bottle, box_1, box_2, and can, as shown on the left. The right side illustrates the process: (#1-#3 - The robot grasps the bottle, #4-#5 - The robot places it into the plastic bin).

Appendix A

Humanoid Policy Appendix

Figure [Figure A.3](#) documents the vertical-grasping setup used for the object-placement generalization study.

Figure [Figure A.4](#) summarizes the pouring task variations used in the autonomous execution evaluation.

A.1 More Ablation Study - Data Collection

Autonomous Whole-body Movement. In Tab. [Table A.1](#), we justify the necessity to minimize body movement in human data collection. Humans tend to move their upper body unconsciously during manipulation (including shoulder and waist movement). However, existing humanoid robots have yet to reach such a level of dexterity. Thus, having these difficult-to-replicate actions in the human demonstrations leads to degraded performance.

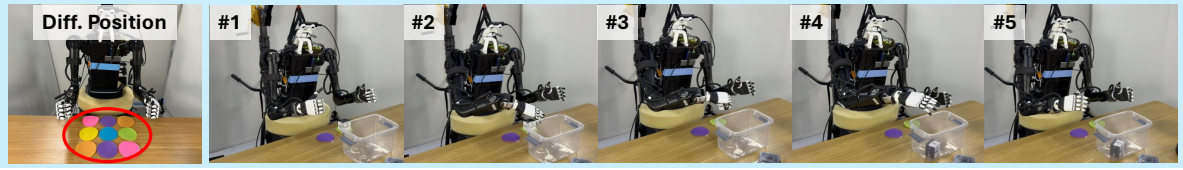


Figure A.3. The robot performs the **vertical grasping** task. As shown on the left, the Dynamixel box is placed in nine different positions for grasping. The right side illustrates the process: (#1-#3 - The robot grasps the box, #4-#5 - The robot places the box into the plastic bin).

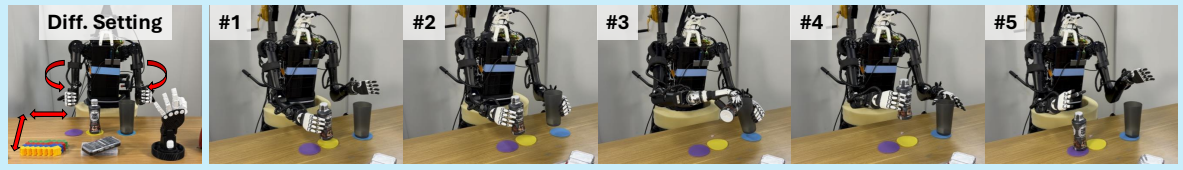


Figure A.4. The robot performs the **pouring** task. The left side shows different settings achieved by varying the robot’s rotation and the table’s position. The right side illustrates the pouring process: (#1 - Right hand grasps the bottle, #2 - Left hand grasps the cup, #3 - Pouring the drink, #4 - Left hand places the cup down, #5 - Right hand places the bottle down).

We hypothesize that such a necessity would be greatly reduced with the development of both whole-body locomotion methods and mechanical designs, but for the currently available platforms, we instruct operators to minimize body movement as much as possible in our dataset.

A.2 In-Depth Analysis of Different Types of Generalization

Human data improves background generalization. We chose to use the *cup passing* task to test background generalization. We prepared four different tablecloths as backgrounds, as shown in Fig. [Figure A.1](#). In terms of training data distribution, the teleoperation data for this task was collected exclusively on the paper background shown in Fig. [Figure A.1](#), whereas the human data includes more than five different backgrounds. This diverse human dataset significantly enhances the generalization ability of the co-trained HAT policy. As shown in Tab. [Table A.3](#), HAT consistently outperforms across all four backgrounds, demonstrating robustness to background variations. In addition, the overall success rate increases by nearly 50% compared to training without human data, highlighting

Table A.1. **Ablation of how human whole-body movement in training demonstrations affects policy rollout.** We collect the same number of demonstrations on the same set of objects for the *grasping* task with or without whole-body movement. Since the robot does not have a natural whole-body movement like humans, it negatively influences the manipulation success rate.

Method	Bottle	Box ₁	Box ₂	Can	Ovr. Succ.
	I.D.	H.D.	H.D.	H.D.	
Without whole-body	8/10	6/10	0/10	7/10	21/40
With whole-body	9/10	3/10	3/10	3/10	18/40

the advantage of utilizing diverse human demonstrations.

Human data improves appearance generalization. To test how co-training improves robustness to perturbations in object textures, we evaluate the *horizontal grasping* policy on novel objects, as shown in Fig. [Figure A.2](#). Specifically, we compare the policy’s performance on the bottle, box₁, box₂, and can, as shown left to right in the first image in Fig. [Figure A.2](#). These objects differ significantly in both color and shape from the bottle used in the teleoperation data distribution.

Since grasping is a relatively simple task, our adjusted policy demonstrates strong learning capabilities even with only 50 teleoperation data samples. The policy can successfully grasp most bottles despite the limited training set. To better highlight the impact of human data, we selected more challenging objects for evaluation. As shown in Tab. [Table A.2](#), human data significantly enhances the policy’s ability to grasp these more difficult objects.

Notably, box₁ appears in the human data, while box₂ does not. Despite this, we observe that co-training with human data still improves overall performance, even on box₂, though its success rate does not increase. This suggests that, beyond direct experience with specific objects, the human data helps the policy learn broader visual priors that enable more proactive and stable grasping behaviors. For box₂, while the success rate remains low partially due to its low height and color similarity to the table, the co-trained HAT policy demonstrates fewer out-of-distribution failures and more actively searches for graspable regions. The failures on box₂ are primarily due to unstable grasping and the small box slipping from the hand, rather than the inability to perceive or locate the object.

Furthermore, adding more human data not only improves performance on objects seen in human training demonstrations (e.g., box₁) but also enhances generalization to completely novel objects (e.g., box₂ and can). We hypothesize that, as the number of objects grows, HAT starts to learn inter-category visual priors that guide it to grasp objects more effectively,

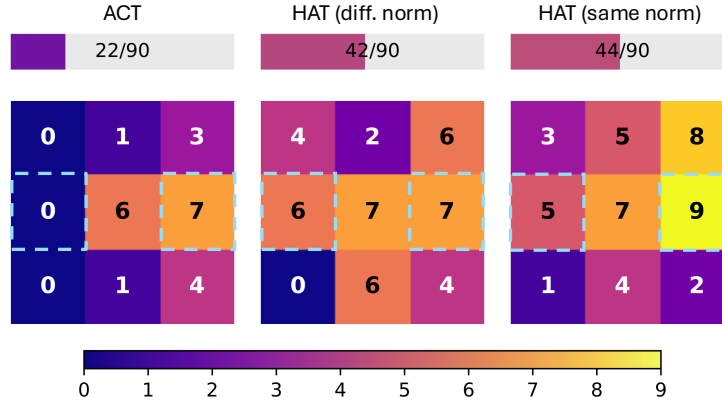


Figure A.5. **Object Placement Generalization.** Performance comparisons of models trained with and without human data on vertical grasping (picking). Each cell in the 3×3 grid represents a $10\text{cm} \times 10\text{cm}$ region where the robot attempts to pick up a box, with numbers indicating successful attempts out of 10. The real-robot data is collected in two cells inside the dashed lines. Notably, our teleoperation data is intentionally imbalanced.

even when they were not explicitly present in the training set.

Human data improves object placement generalization. Finally, we introduce variations in object placements that are not present in the real-robot training demonstrations and specifically investigate this in the *vertical grasping (picking)* task. In this task, we intentionally constrain the robot data collection to object placements within a subset of cells, while human vertical grasping data covers a much more diverse range of settings.

To systematically analyze the impact of human data, we evaluate model performance on a structured 3×3 grid, where each cell represents a $10\text{cm} \times 10\text{cm}$ region for grasping attempts. The numbers in each cell indicate the number of successful picks out of 10 trials. Real-robot training data is collected from only two specific cells, highlighted with dashed lines.

A key detail in our teleoperation data distribution is that 50 picking attempts are collected from the right-hand side grid and only 10 from the left-hand side grid. This imbalance explains why policies trained purely on teleoperation data struggle to grasp objects in the left-side grid. We observe that models trained solely on robot data fail to generalize to unseen cells, whereas cross-embodiment learning with human data significantly improves generalization, doubling the overall success rate. Figure [Figure A.5](#) reports the corresponding placement-generalization results.

Table A.2. **Object Appearance Generalization:** In the horizontal grasping task, we evaluated the grasping performance by attempting to grasp each object 10 times and recorded the success rate.

Method	Bottle	Box ₁	Box ₂	Can	Ovr. Succ.
	I.D.	H.D.	O.O.D.	O.O.D.	
ACT	8/10	5/10	1/10	1/10	16/40
HAT	8/10	7/10	1/10	4/10	21/40

Table A.3. **Background Generalization:** In the cup passing task, we evaluate the passing performance by recording the number of failures or retries needed to complete 20 cup-passing trials.

Method	Paper	Wooden	Red	Green	Ovr. Succ.
	I.D.	H.D.	O.O.D.	O.O.D.	
ACT	19/20	14/20	12/20	10/20	55/80
HAT	20/20	16/20	18/20	18/20	72/80

A.3 In-Depth Comparison between Humanoid A and Humanoid B Configurations

This section presents a detailed comparison of the two humanoid platforms, referred to as Humanoid A and Humanoid B, with a focus on joint structure and implications for manipulation capabilities. We restrict our analysis to the arm configurations, as other parts of the body were not exclusively explored in this work.

While morphologically similar, these two humanoids have drastically different arm configurations that create hurdles in direct policy transfer. Besides differences in motor technical specs such as torque and types of encoder (Humanoid B has absolute motor position encoders), they also have different mechanical limits. The range of motion (ROM) for the first four proximal joints, shoulder_pitch, shoulder_roll, shoulder_yaw, and elbow, differs across the two platforms. Humanoid B exhibits a consistently wider ROM, which allows a wider set of reachable configurations and increases the manipulability of the arm in constrained environments. Table [Table A.4](#) summarizes the ROM values for these shared joints.

A more significant architectural divergence is observed at the wrist. Humanoid A includes a single distal joint, wrist_roll, providing limited wrist articulation. This restricts end-effector dexterity and constrains in-hand manipulation strategies to a single rotational degree of freedom. In contrast, Humanoid B is equipped with a complete wrist mechanism composed of

three independently actuated joints: wrist_pitch, wrist_roll, and wrist_yaw. These additional degrees of freedom allow for full orientation control of the end-effector, enabling tasks that require precise alignment, rotation, and fine adjustment of object poses.

Table A.4. Joint Range of Motion Comparison between Humanoid A and B (in degrees)

Joint	Humanoid A	Humanoid B
shoulder_pitch	-164° to $+164^{\circ}$	-180° to $+90^{\circ}$
shoulder_roll	-19° to $+178^{\circ}$	-21° to $+194^{\circ}$
shoulder_yaw	-74° to $+255^{\circ}$	-152° to $+172^{\circ}$
elbow	-71° to 150°	-54° to 182°
wrist_roll	-175° to 175°	-172° to 157°

Appendix B

EmbodHIO Project Appendix

This appendix collects the implementation details that support Chapter 3. The main chapter focuses on the method and headline experiments; here we spell out the guidance formulation, retargeting and RL setup, the policy inputs, rewards, termination conditions, hyperparameters, and the extra qualitative and quantitative results.

B.1 Flow Matching and Energy Guidance

MAD-HOI [66] generates continuous motion latents using a flow-matching objective. Rather than learning a sequence of discrete denoising transitions, the model learns a time-dependent velocity field $\mathbf{v}_\theta(\mathbf{x}, t, \mathbf{z})$ that transports samples from a simple noise distribution toward the data distribution conditioned on context $\mathbf{z}$. Under a linear interpolation path between clean latents and Gaussian noise, the learned vector field defines a probability-flow ODE,

$$\dot{\mathbf{x}} = \mathbf{v}_\theta(\mathbf{x}, t, \mathbf{z}), \quad (\text{B.1})$$

whose integration progressively transforms noisy latent states into valid motion samples. Following classifier-style guidance in diffusion and flow-based generative models, we bias sampling with an auxiliary energy function $E(\mathbf{x})$,

$$\dot{\mathbf{x}} = \mathbf{v}_\theta(\mathbf{x}, t, \mathbf{z}) - w(t)\nabla_{\mathbf{x}}E(\mathbf{x}), \quad (\text{B.2})$$

where $w(t) \geq 0$ controls the guidance strength. The gradient term steers trajectories toward lower-energy regions while preserving the learned prior. In EmbodHIO, $E(\mathbf{x})$ corresponds to the inverse-kinematics residual used during inference.

ODE integration. During sampling, we integrate the guided probability-flow ODE with a fourth-order Runge–Kutta solver using 20 integration steps for guided generations.

B.2 Policy Observations

Table Table B.1 makes the policy input explicit. The policy is not trained from a single instantaneous state, but from a short history of robot proprioception, previous actions, future

Table B.1. Policy observations for bimanual manipulation on articulated objects.

Observation	Interpretation
$\{\mathbf{q}_{t-k}^{\text{robot}}\}_{k=0}^8$	Robot joint positions over the past 9 steps (proprioception).
$\{\dot{\mathbf{q}}_{t-k}^{\text{robot}}\}_{k=0}^8$	Robot joint velocities over the past 9 steps.
$\{\mathbf{a}_{t-k}\}_{k=0}^8$	Previous actions over the past 9 steps ($n_a=39$ controlled DOFs).
ϕ_t	Normalized episode phase $2t/T_{\max} - 1$.
$\hat{\mathbf{q}}_{t+\Delta}^{\text{robot}} - \mathbf{q}_t^{\text{robot}}, \Delta \in \{1, 2, 4, 8, 16\}$	Future reference joint offsets.
$T_{t,k}^{\text{obj}}, R_{t,k}^{\text{obj}}, k \in \{1, 2\}$	3D position and 4D orientation of both articulated object parts.
α_t^{obj}	1D object articulation (joint opening).
$d_p, d_r, (\alpha_{t+h}^{\text{obj}} - \alpha_t^{\text{obj}}), h \in \{1, 2, 4, 8, 16\}$	Future tracking errors vs. reference object pose.
$\hat{\mathbf{p}}_{t+h}^{\text{obj}}, h \in \{1, 2, 4, 8, 16\}$	Reference object root positions at future timesteps.
$\hat{\mathbf{p}}_t^{\text{obj}} - \mathbf{p}_t^{\text{obj}}$	Current object root position error w.r.t. reference.

robot references, and future object references.

The observation structure matters because it gives the LSTM temporal context for both tracking and anticipation.

B.3 IK Optimization Details

IK objective. The guidance energy is computed from a differentiable inverse-kinematics optimization on the target humanoid embodiment. The optimization combines a primary wrist-pose tracking objective with regularization terms that improve kinematic feasibility and base consistency. Table [Table B.2](#) summarizes the objective components and their weights. After solving the primary IK objective, we apply a secondary null-space optimization to improve the resulting arm posture.

Null-space optimization. During guidance, wrist-pose tracking defines the primary IK target, while the remaining terms regularize the solution toward stable and feasible arm and base configurations. In particular, the object-clearance term discourages arm configurations that pass too close to the manipulated object. The primary optimization is solved with Levenberg–Marquardt. We then apply the null-space manipulation objective to adjust the resulting arm configuration without altering the target wrist pose, yielding postures that are better suited for downstream retargeting and policy learning.

Optimization hyperparameters. Table [Table B.3](#) lists the Levenberg–Marquardt settings used throughout the experiments.

Solver settings. The same optimization settings are used across all embodiments and tasks. We use Levenberg–Marquardt with a maximum of 10 iterations per timestep, a gradient-norm

Table B.2. Components of the IK guidance optimization objective. Wrist tracking acts as the primary task objective, while the remaining terms regularize the solution toward kinematically feasible arm and base configurations. A secondary null-space optimization is applied after solving the primary IK problem.

Component	Weight	Description
Wrist position tracking	$w_{\text{pos}} = 250$	Tracks target wrist position.
Wrist orientation tracking	$w_{\text{ori}} = 5$	Tracks target wrist orientation.
Rest pose regularization	$\lambda_{\text{rest}} = 10$	Encourages solutions near a nominal arm configuration.
Joint-limit penalty	$w_{\text{joint}} = 100$	Penalizes configurations outside joint limits.
Configuration smoothness	$w_{\text{smooth}} = 1$	Penalizes deviation from the previous IK solution.
Root translation prior	$w_{\text{root}} = 100$	Encourages consistency with the preferred robot base position.
Object clearance	$w_{\text{clear}} = 100$	Penalizes configurations violating a minimum clearance radius around the object.
Null-space manipulation	$\lambda_{\text{manip}} = 0.03$	Secondary optimization preserving wrist pose while improving arm posture.

termination threshold of 10^{-2} , and implicit differentiation to propagate guidance gradients through the IK solve. This was sufficient to obtain stable wrist-pose solutions while keeping guidance runtime practical during sampling.

B.4 Additional Singularity Analysis on Unitree H1-2

Figure [Figure B.1](#) reports the minimum Jacobian singular value and maximum condition number for Unitree H1-2 using the same evaluation protocol as the main chapter. Compared to Unitree G1, the trends are less pronounced, suggesting that the effect of IK guidance depends on the target embodiment and task distribution. Nevertheless, guidance changes the kinematic profile of the generated motions while preserving the ability to transfer the same HOI generator across substantially different humanoid morphologies.

B.5 Reward and Termination Design

Tables [Table B.4](#) and [Table B.5](#) separate the task term, imitation terms, stabilization penalties, and failure conditions.

Table B.3. Optimization parameters used by the IK guidance module.

Parameter	Value
LM iterations	10
LM damping coefficient	1e-6
Implicit differentiation	Enabled

H1-2 — unguided active $sv_{\min} < 0.0001$

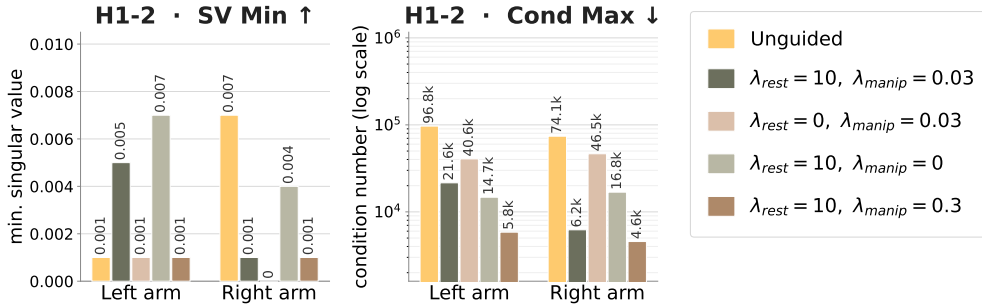


Figure B.1. Minimum singular value and maximum condition number for Unitree H1-2. Higher singular values and lower condition numbers indicate better kinematic conditioning.

The reward split is the same decomposition used in Chapter 3: upper-body control tracks reachability, finger control tracks contact, and contact terms preserve the object interaction once the arm is in place.

B.6 Training Hyperparameters

Table Table B.6 lists the PPO setup used in this method.

B.7 Computational Cost

Table Table B.7 summarizes the approximate computational cost of the major stages in our pipeline. IK guidance is applied only during inference and therefore introduces no additional training cost for the generative model. While guidance increases generation time relative to standard MAD-HOI sampling, it remains substantially less expensive than retraining a robot-specific motion generator for each embodiment.

All runtimes except PPO training were measured on a workstation equipped with an NVIDIA A100 GPU. PPO training was measured on an RTX 4090 GPU. All runtimes are intended only as approximate references and can vary from task to task.

Table B.4. Reward terms used for task execution and motion stabilization.

Reward Term	Interpretation
$r_{\text{task}} = \left\ T_t^{\text{obj}} - \hat{T}_t^{\text{obj}} \right\ _2 \cdot d_{\text{rot}}(R_t^{\text{obj}}, \hat{R}_t^{\text{obj}}) - \beta_a \left(\alpha_t^{\text{obj}} - \hat{\alpha}_t^{\text{obj}} \right)^2$	Rewards matching reference object position, orientation, and articulation in world frame.
$r_{\text{imi,upper}} = \left\ \mathbf{q}_t^{\text{upper}} - \hat{\mathbf{q}}_t^{\text{upper}} \right\ _2^2$	Rewards tracking of arms and torso joints to reference motion.
$r_{\text{imi,dex}} = \left\ \mathbf{q}_t^{\text{dex}} - \hat{\mathbf{q}}_t^{\text{dex}} \right\ _2^2$	Rewards tracking of finger joints to reference motion.
$r_{\text{contact}} = \left\ \mathbf{c}_t - \hat{\mathbf{c}}_t \right\ _2$	Rewards matching simulated contact locations to reference contact locations on hand links.
$r_{\text{act}} = \left\ \mathbf{a}_t \right\ _2$	Penalizes large actions.
$r_{\text{act-rate}} = \left\ \mathbf{a}_t - \mathbf{a}_{t-1} \right\ _2^2$	Penalizes large action rates.
$r_{\text{ctrl}} = \left\ \boldsymbol{\tau}_t \right\ _1$	Penalizes large joint control torques.
$r_{\text{vel}} = \left\ \dot{\mathbf{q}}_t \right\ _1$	Penalizes large joint velocities.
$r_{\text{acc}} = \left\ \ddot{\mathbf{q}}_t \right\ _1$	Penalizes large joint accelerations.

Table B.5. Episode termination conditions during training.

Termination Condition	Interpretation
$t \geq T_{\text{max}}$	Ends episode after reaching maximum horizon.
NaN in simulation state	Ends episode on numerical failure in robot or object state.
$\max_i \left\ \mathbf{q}_t^{\text{upper}} - \hat{\mathbf{q}}_t^{\text{upper}} \right\ _2^2 > \epsilon_{\text{upper}}$	Ends episode when arm or torso joints deviate too far from reference motion.
$\left\ \mathbf{q}_t^{\text{dex}} - \hat{\mathbf{q}}_t^{\text{dex}} \right\ _2^2 > \epsilon_{\text{dex}}$	Ends episode when finger joints deviate too far from reference motion.
$p_{t,z}^{\text{obj}} < h_{\text{table}}$	Ends episode when the object falls below the table.
$\left(\alpha_t^{\text{obj}} - \hat{\alpha}_t^{\text{obj}} \right)^2 > \epsilon_a$	Ends episode when object articulation deviates too far from the reference trajectory.

B.8 Additional Results

We also include additional qualitative visualizations of generated and executed manipulation behaviors on the Unitree G1 platform (Figure [Figure B.2](#)) and a qualitative comparison between IK-guided and unguided HOI generations (Figure [Figure B.3](#)), illustrating the improved kinematic feasibility achieved by our guidance method while preserving the semantic intent of the interaction.

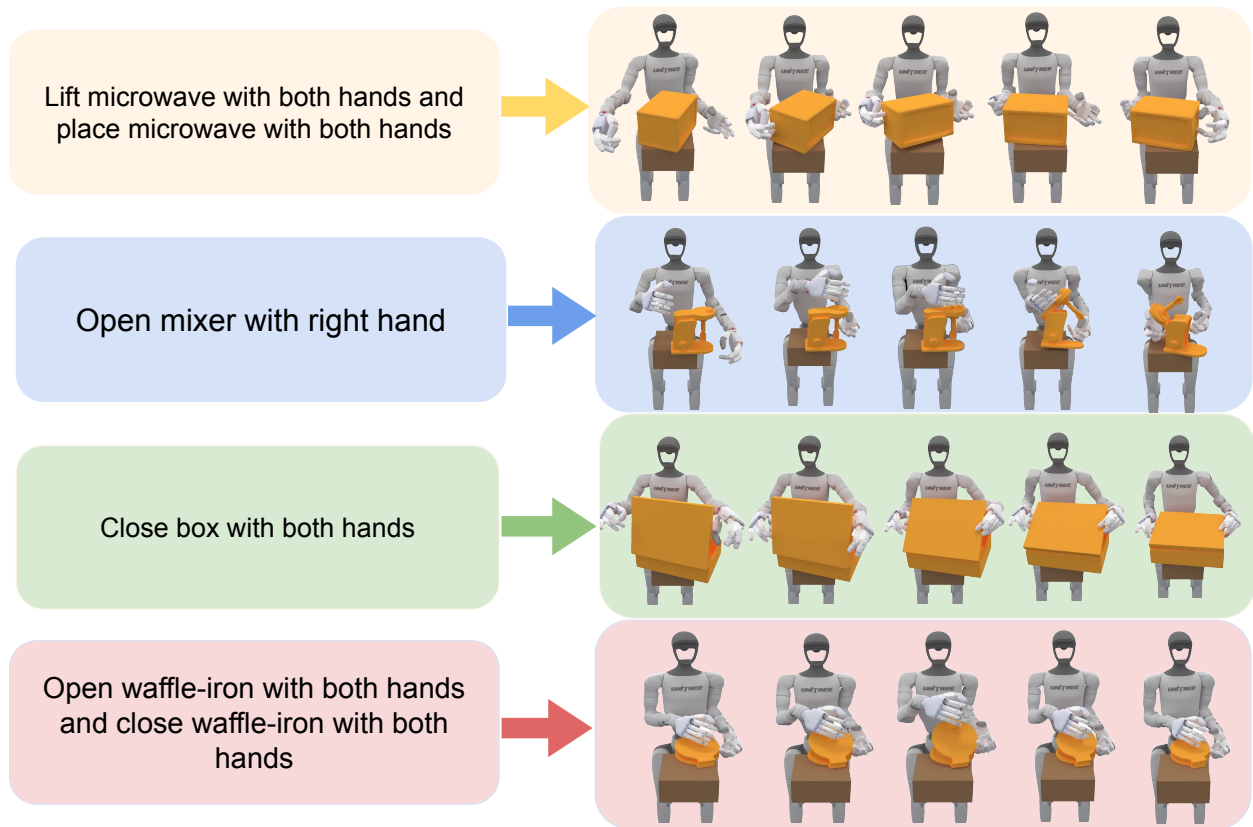


Figure B.2. Additional qualitative manipulation tasks executed on the Unitree G1 humanoid using policies trained from EmbodyHOI-generated trajectories. The examples demonstrate the diversity of articulated-object interactions supported by our pipeline, including opening, closing, lifting, and bimanual manipulation behaviors.

Table B.6. Hyperparameters used in the EmbodyHOI RL experiments.

Category	Parameter	Value
<i>PPO / training</i>		
	Parallel environments B	8096
	Training epochs	5000
	Rollout horizon (steps per env)	16
	Learning rate	3×10^{-4}
	Discount γ / GAE τ	0.99 / 0.95
	PPO clip ϵ	0.2
	Entropy coefficient	0.0055
	Bounds loss coefficient	10^{-4}
	Minibatch size	$8B$
	PPO epochs per rollout	$\max(1, \lfloor 5B/4096 \rfloor)$
	Reward shaper scale	0.1
	Policy MLP	512, 512, 256, 128 (ELU)
<i>Environment</i>		
	Episode length $T_{\max}$	Task dependent
	Object demo horizons	$\{1, 2, 4, 8, 16\}$
	Robot history steps	$\{0, \dots, 8\}$
	Robot future steps	$\{1, 2, 4, 8, 16\}$
<i>Reward weights</i>		
	w_{task}	1.3
	w_{bc} (upper body)	0.9
	w_{finger}	0.15
	w_{contact}	0.7
<i>Reward shaping</i>		
	$\beta_p, \beta_r, \beta_a$	10, 1, 5
	β_{bc} (upper)	30
	β_{finger}	20
	β_{contact}	30
<i>Penalties</i>		
	λ_{act}	0.02
	λ_{rate}	0.003
	λ_{vel}	0.03
	λ_{acc}	0.025
	λ_{ctrl}	0.01
<i>Termination thresholds</i>		
	ϵ_{upper}	0.12
	ϵ_{dex}	0.2
	ϵ_a (articulation)	0.4

Table B.7. Approximate runtime of major pipeline stages.

Stage	Runtime
MAD-HOI generation	5 s
IK-guided generation	300 s
PPO training	4.0 h

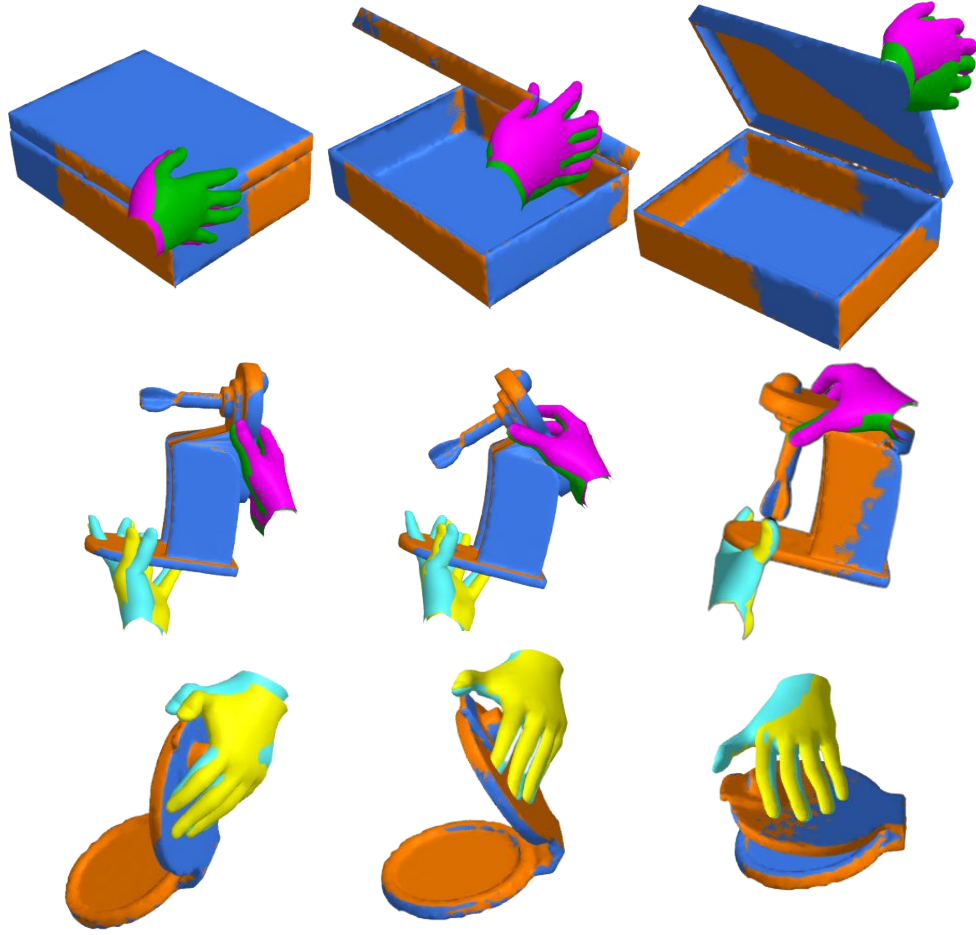


Figure B.3. Qualitative comparison of unguided (left:cyan; right:green) MAD-HOI generations and IK-guided (left:yellow; right:magenta) generations. The guidance objective steers wrist trajectories toward kinematically feasible configurations for the target humanoid embodiment while largely preserving the semantic structure of the generated hand-object interaction.