Tool-as-Interface: Learning Robot Policies from Human Tool Usage through Imitation Learning

Haonan Chen[1], Cheng Zhu[1], Yunzhu Li[2], Katherine Driggs-Campbell[1]
[1] University of Illinois, Urbana-Champaign [2] Columbia University
https://tool-as-interface.github.io

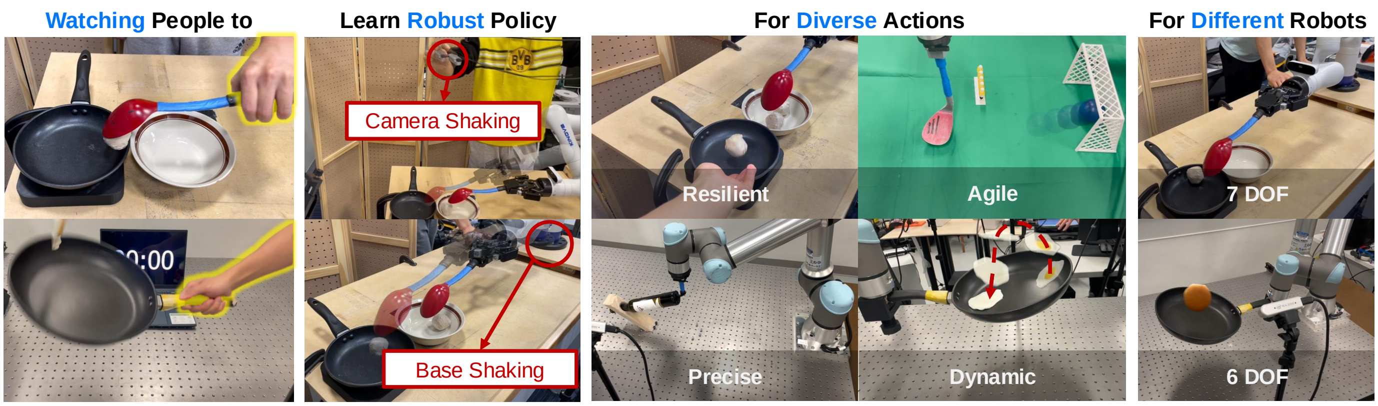


Fig. 1: **Tool-as-Interface.** We propose a scalable data collection and policy learning framework designed to transfer diverse, intuitive, and natural human data into effective visuomotor policies. The framework enables robots to learn robust policies that can operate effectively under challenging conditions, such as base and camera movement, and achieve high performance on a variety of complex manipulation tasks.

***Abstract*—Tool use is critical for enabling robots to perform complex real-world tasks, and leveraging human tool-use data can be instrumental for teaching robots. However, existing data collection methods like teleoperation are slow, prone to control delays, and unsuitable for dynamic tasks. In contrast, human natural data—where humans directly perform tasks with tools—offers natural, unstructured interactions that are both efficient and easy to collect. Building on the insight that humans and robots can share the same tools, we propose a framework to transfer tool-use knowledge from human data to robots. Using two RGB cameras, our method generates 3D reconstruction, applies Gaussian splatting for novel view augmentation, employs segmentation models to extract embodiment-agnostic observations, and leverages task-space tool-action representations to train visuomotor policies. We validate our approach on diverse real-world tasks, including meatball scooping, pan flipping, wine bottle balancing, and other complex tasks. Our method achieves a 71% higher average success rate compared to diffusion policies trained with teleoperation data and reduces data collection time by 77%, with some tasks solvable only by our framework. Compared to hand-held gripper, UMI [16], our method cuts data collection time by 41%. Additionally, our method bridges the embodiment gap, improves robustness to variations in camera viewpoints and robot configurations, and generalizes effectively across objects and spatial setups.**

I. Introduction

Tool use is essential to how humans interact with and transform their environment [67, 41]. For instance, humans use a pan to fry food and flip it, ensuring even cooking on both sides. Despite its significance, tool use beyond parallel jaw grippers remains underexplored in robotics, with research primarily focused on simpler tasks like grasping and pick-and-place operations [6, 46, 10, 36, 35]. In this paper, we focus on cost-effective data collection and efficient training of robot policies to rapidly acquire tool-use skills.

Imitation learning provides a promising pathway for robots to acquire tool-use skills by directly learning from human demonstrations [21, 26, 27, 29]. The paradigm excels in handling diverse tool-use tasks, as it bypasses the need for task-specific programming by relying solely on human demonstrations. However, its full potential hinges on addressing key challenges in collecting high-quality training data. Various teleoperation systems [33, 56, 8, 20, 29, 50, 84, 68, 39, 66, 45, 9] and hand-held grippers [65, 19, 53, 47, 43] have been developed to facilitate the collection of high-quality data. Teleoperation methods, such as kinematic replication and hand or body retargeting, show great potential [85, 22, 74, 62, 48]. However, their reliance on direct access to robot hardware limits both practicality and scalability. Hand-held grippers [57, 16] offer an alternative by enabling demonstrations in diverse environments. While they reduce dependency on robotic systems, their high costs and the technical expertise required for tasks like 3D printing and assembly restrict their accessibility to a specialized group of users.

To address these limitations, we turn to human interaction data — a natural, intuitive method through which humans interact with their environment during everyday activities without relying on external devices or specialized setups. Human interaction data refers to the natural process in which humans use their hands to operate tools and interact with and manipulate the environment freely. Unlike controlled demonstrations that require expensive hardware or meticulous preparation, human interaction data involves the spontaneous use of tools to interact with the environment. Human interaction data is an accessible, scalable, and cost-effective approach to data collection, requiring no prior knowledge or technical expertise, such as 3D printing or assembly, from participants. However, existing methods struggle to fully harness the potential of human interaction data. Key challenges include the embodiment gap and the reliance on single-view data, which limits the insights that can be drawn from interaction data [60, 4, 38, 5, 64, 42, 55].

Our framework addresses these challenges by leveraging the observation that humans and robots can share the same tools. We propose a novel approach that utilizes human interaction data to train robust and adaptable robot policies for diverse tool-use tasks (Figure 1). Our method minimizes reliance on expensive hardware, making data collection more scalable and accessible to non-experts. By capturing 3D information using two RGB cameras and generating 3D reconstructions, our method enables view-invariant policy learning through novel view augmentation. To facilitate direct policy transfer from human interaction data to robotic systems, we employ a segmentation model to filter out embodiment-specific information. Additionally, we leverage task-space tool-action representations to ensure robustness to variations in robot base configurations.

Our contributions are as follows:

1) We propose a novel framework that leverages two-view human interaction data to enable scalable, intuitive, and cost-effective data collection for training robot policies on complex tool-use tasks.
2) We validate the effectiveness of our approach on diverse challenging real-world tool-use tasks (e.g., nail hammering, meatball scooping, pan flipping, wine bottle balancing, and soccer ball kicking). Our method achieves a 71% higher success and 77% less data collection time than diffusion policies trained on SpaceMouse [17] or Gello [74] data, with some tasks solvable only by our method. Our method also outperforms handheld grippers like UMI [16], reducing collection time by 41%.
3) We provide an extensive analysis of our method's robustness under varying conditions, including changes in camera poses, robot base movements, and human-induced perturbations. Additionally, we conduct ablation studies to evaluate the effects of different design choices on policy performance, including embodiment segmentation, random cropping, and novel view augmentation.

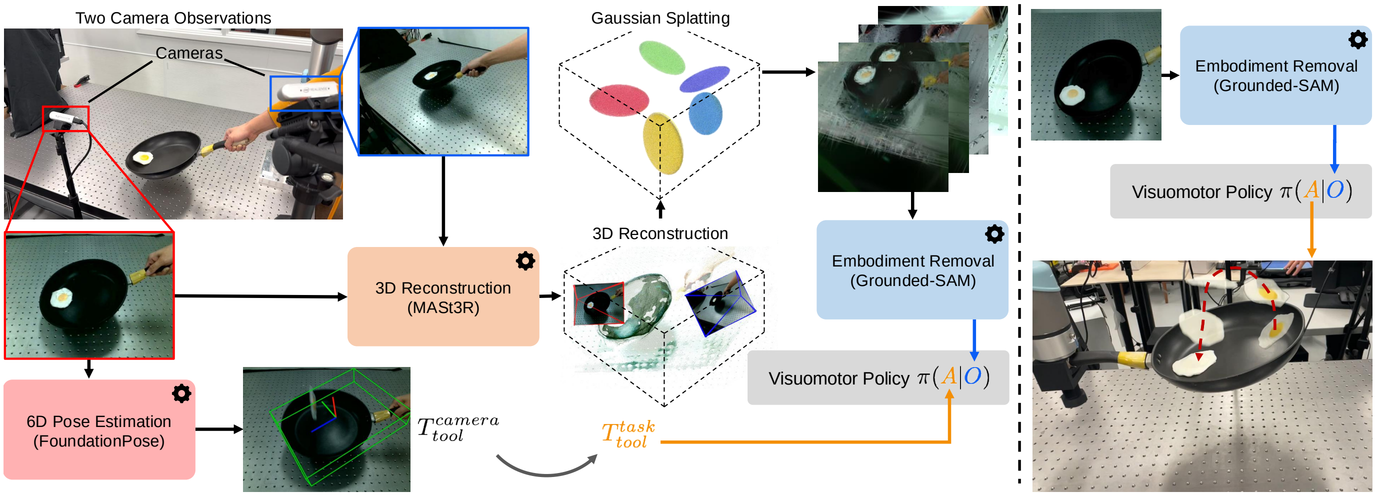
II. Related Works

A. Data Collection for Robot Learning

High-quality data is essential for training robots to learn and generalize across tasks. Simulation data has been widely used for its cost-effectiveness and scalability [86, 44, 28, 81], but the gap between simulation and real-world performance remains a persistent challenge. To overcome this, many researchers have turned to real-world teleoperated demonstrations, which offer more reliable transferability by minimizing the domain gap between training and testing environments [88, 14, 15, 56, 73]. Advancements in leader-follower devices, like ALOHA [85, 22] and GELLO [74], have simplified robot demonstration data collection but remain tied to specific robot platforms. More recently, portable tools such as hand-held grippers, e.g., UMI [16] and LEGATO [57], have emerged as a promising alternative for in-the-wild data collection. Yet, their high cost and the need for custom robot modifications continue to limit widespread adoption. Tool-based policy representations have emerged as an effective way to collect data for robot learning. MimicTouch uses tactile-based tools for contact-rich manipulation [79], while ScrewMimic models bimanual tasks as constrained screw motions for learning from human videos [3]. However, tactile methods require extra hardware, and the screw motion assumption may not hold. Another line of work by Wen et al. [70] aims to learn category-level representations from a single demo to transfer pose trajectories across similar objects [70], but assumes a static target (e.g., battery slot), limiting real-world applicability in real-world tasks where the spatial configuration of objects may change. Unlike these methods, our method only requires natural human data, without assuming access to tactile sensors or requiring constrained screw motion models, making it significantly more cost-effective, scalable, and accessible.

B. Cross-Embodiment Policy Learning

Cross-embodiment policy learning enables robots to transfer policies across embodiments, such as those with different kinematic structures [25, 52, 77]. Prior work has explored conditioning policies on embodiment representations using multi-embodiment datasets [18, 69, 80, 13, 24, 52, 2], but these approaches often face challenges in effectively leveraging human interaction data. Recent approaches utilize human data, such as estimating point flow from human video [72] or generating latent plan or high-level plans from human data [37, 68]. However, their dependence on robot data for low-level control limits scalability when robot data is expensive or difficult to collect. Additionally, prior works highlight the importance of masking human and robot embodiments for visual consistency [4, 31]. However, Bahl et al. [4] relies on predefined motion primitives, while Kareer et al. [31] still requires robot data with human data as augmentation. Our approach adopts a similar masking idea but enables robots to learn freely, even agile motions from human videos, without any robot data. Another line of work tokenizes observation inputs and action outputs into a unified transformer network, enabling generalized policy learning across embodiments [59, 76]. However, these methods require large models and extensive datasets, making them resource-intensive and time-consuming to train, and lack the capability for direct policy transfer between embodiments. Overall, reliance on robot-specific hardware and data restricts scalability and accessibility. In contrast, our method leverages natural human data, eliminating the need for robot data as a training source.



(a) Data Collection and Policy Training (b) Policy Evaluation

Fig. 2: **Policy Design.** Human interaction data was collected using two RGB cameras and processed through the foundation model MASt3R [34] to generate 3D reconstructions. Using 3D Gaussian splatting, we sampled novel views to augument the dataset. The human hand (embodiment) was segmented out from the images to create embodiment-agnostic observations, which serve as inputs to the policy. To label actions for policy training, a pose estimation model (FoundationPose [71] in this work) was used to extract the tool's pose in the camera frame, $T_{\text{tool}}^{\text{camera}}$. A coordinate transformation was then applied to compute the tool's pose in the task space, $T_{\text{tool}}^{\text{task}}$. Finally, a diffusion policy was implemented as the visuomotor policy to enable effective learning and execution.

C. Cross-Viewpoint Policy Learning

Viewpoints can change as robots manipulate or interact with their environment. Existing approaches address viewpoint changes by learning view-invariant latent representations [58, 11], leveraging equivariant 3D representations [63, 23, 89], or augmenting input data with diverse viewpoints [51, 87]. An early attempt by Sharma et al. [61] introduced a hierarchical model that generates first-person sub-goals from third-person demonstrations via a high-level module, while a low-level embodiment-specific module predicts actions. Recent work by Chen et al. [12] leveraged the SAM model [32] and ControlNet [83] to transform the viewpoint of one robot into another, performing view synthesis with ZeroNVS [54] to augment datasets and enhance policy learning across robots and viewpoints. Another approach by Yuan et al. [82] combined reinforcement learning with multi-view representation learning and a Spatial Transformer Network (STN) to improve policy robustness in visually complex environments. However, these methods face limitations such as scalability issues in hierarchical models, the complexity of controlled image editing, high computational costs in view synthesis, and the need to build sophisticated simulators for reinforcement learning when adapting to diverse environments. Compared to previous work, our method enables efficient and scalable cross-viewpoint transfer from human data by leveraging two RGB images with the MASt3R [34] foundation model for 3D reconstruction and employing Gaussian splatting for fast and cost-effective novel view synthesis.

III. Problem Statement

We formulate the robotic manipulation task as a Markov Decision Process, where the goal is to learn a policy $\pi : \mathcal{O}^r \rightarrow \mathcal{A}$ that enables a robot to perform a given task. The robot's observation space, $\mathcal{O}^r$, consists of a single-view RGB image $I^r \in \mathbb{R}^{128\times128\times3}$ and proprioception data $x^r \in SE(3)$. To train the policy, we use an imitation dataset composed of N human play, denoted as $D = \{(O_0^h, O_1^h, \dots)\}_{n=1}^N$. Each human play observation $\mathcal{O}^h$ consists of two RGB images captured from different viewpoints: $\mathcal{O}^h = \{I^{v1}, I^{v2}\}$, where $I^{vi} \in \mathbb{R}^{480\times640\times3}$. We preprocess the dataset to get the action by using a 6D pose estimation and tracking model, resulting in $D = \{(O_0^h, a_0, O_1^h, a_1, \dots)\}_{n=1}^N$, where each action $a \in SE(3)$.

Our objective is to train a robot to perform the same task as demonstrated in the human dataset while overcoming the embodiment differences. We assume that the tool is rigid, the tool and the end effector are rigidly attached during deployment, and the transformation between the tool and the end effector is estimated once before deployment. Our approach aims to allows the robot generalize task execution across embodiments while retaining the core skills demonstrated by humans.

IV. Our Robot Policy Learning Framework

Our framework enables the direct transfer of human play data into deployable robot policies. It is designed to fulfill the following key objectives:

- **Support for Dynamic and High-Precision Tasks:** Human play, with its inherent fluidity, enables the execution of highly dynamic tasks. Examples include flipping an egg in a pan or performing other actions that require swift, accurate, and natural motions — challenges that are often difficult to address with traditional teleoperation systems or handheld grippers.
- **Robustness:** The framework ensures robust performance under dynamic conditions, enabling reliable task execution even with moving or shaking cameras. While broader deployment on mobile platforms such as quadrupeds or humanoids remains an open challenge, our design and

experimental results suggest strong potential for generalization to dynamic environments.
- **Generalization Across Robotic Embodiments and Object Categories:** The framework demonstrates broad generalizability, validated on robotic platforms such as the UR5e and Kinova Gen3. It extends its capabilities to manipulate a wide range of object categories, showcasing its adaptability to various tasks, setups, and environments.
- **Affordability and Accessibility:** The framework requires only two monocular RGB cameras, such as smartphones, webcams, or RealSense cameras. With approximately 7.14 billion smartphones worldwide — covering around 90% of the global population — this setup is accessible to almost anyone [30]. By relying solely on RGB cameras, the framework eliminates the need for designing, printing, or manufacturing additional hardware during the data collection, ensuring a cost-effective and inclusive solution.
- **Intuitive and Natural Interaction:** Users can interact naturally, without the need for specialized equipment or additional tools. Using their bare hands and common tools, participants can intuitively perform a variety of tasks. Our approach removes technical barriers associated with 3D printing and other hardware setups, fostering a seamless, user-friendly experience for data collection.

The following sections outline the framework's design (Figure 2), its underlying principles, and logistical considerations for development and deployment.

A. Tool Usage for Data Collection and Manipulation

Humans naturally and intuitively use tools for everyday tasks such as cooking, eating, cleaning, and interacting with the world. Tools act as extensions of human actions, enabling diverse interaction with objects. The natural relationship between tools and objects provides an ideal interface for training robots to mimic human actions using tools, with minimal gap between human and robot tool usage. While grasping and pick-and-place tasks have been extensively studied in previous works, our work focuses on enabling robots to use the same tools humans commonly employ to interact with their environment effectively with the following benefits:

- **Minimized Embodiment Gap:** Abstracting actions to the tool pose reduces morphological dependency, enabling policies to generalize across embodiments.
- **Scalable Data Collection:** Simplifies the data collection process by eliminating the need for costly robot-specific demonstrations, making our method more accessible.

During data collection, humans can naturally use tools with their hands without requiring additional devices. For robot deployment, the tool can be attached to the robot in two ways:

- **Rigid Grasping:** Grasping or picking, which has been extensively studied in prior works, is demonstrated in our Kinova Gen3 robot experiments and involves the robot securely grasping the tool.
- **Customized Fast Tool Changer:** Designed for versatility, the tool changer is compatible with any robot that uses the ISO 9409-1-50-4-M6 flange, as demonstrated in our UR5e experiments.

B. Embodiment-Agnostic Perception

To encourage cross-embodiment transfer, we adopt a strategy that reduces the perception gap between the training and deployment phases. During training, human play data is collected, featuring human hands interacting with tools and objects. In deployment, robots execute the learned tasks. As showcased in our experiments, the visual differences between human hands and robotic end-effectors can introduce discrepancies that hinder generalization. To address this, we employ Grounded-SAM [49] to segment and mask out the embodiments in each phase. During training, human hands are masked, while during deployment, the robotic embodiments are masked, which ensures that the remaining parts of the scene in both training and testing phases appear visually similar. By aligning perception across embodiments, the framework mitigates distractions caused by embodiment-specific features, enabling better generalization to human-to-robot policy transfer.

C. View Augmentation

We use cameras for data collection due to their availability. With approximately 7.14 billion smartphones equipped with cameras, our approach can scale effectively [30]. However, using data from a single camera introduces challenges such as a lack of 3D perception and sensitivity to camera pose.

a) 3D Reconstruction: To address these issues, we use MASt3R [34], an image-matching model that reconstructs accurate 3D environments from two RGB images. This eliminates the need for additional depth sensors, which are less common and consume more power compared to RGB sensors. We use two cameras, which is the minimum requirement for 3D reconstruction, to capture demonstration data. Using only one camera can lead to scale ambiguity in monocular settings, where poses can be scaled by an arbitrary, scene-dependent factor. By capturing images from two viewpoints, MASt3R reconstructs a 3D point cloud without requiring camera extrinsics or intrinsics, and globally aligns point maps within a multi-view 3D reconstruction framework. The process results in high-quality 3D representations.

b) Data Augmentation: Using 3D Gaussian splatting, we model the scene and synthesize novel viewpoints from human interaction data, which effectively augments the dataset, generating additional perspectives even if the training data was captured from only two views. These synthetic viewpoints provide the robot with a multi-angle understanding of the scene, allowing the policy to be trained on a more diverse and comprehensive set of visual inputs. Additionally, we apply random cropping to the images for data augmentation before feeding them into the policy network, following the approach from diffusion policy [14, 15]. Random cropping further improves the method's robustness, enabling the policy to generalize better to variations in visual inputs.

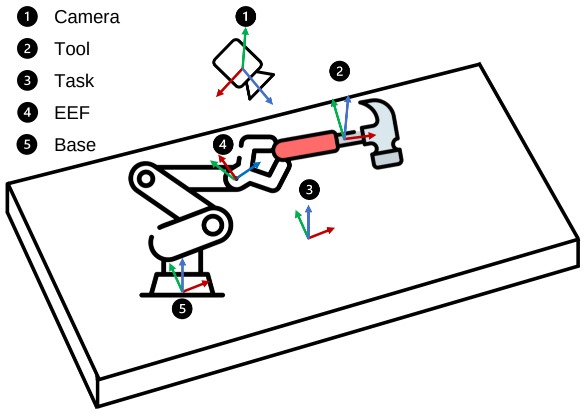


Fig. 3: **Coordinate System Diagram.** This diagram illustrates the Camera frame, Tool frame, Task space frame, End-Effector (EEF) frame, and Base frame. We use $T_{\text{tool}}^{\text{task}}$ as the action representation for the visuomotor policy output.

D. Action Representation for Tool Manipulation

To support general tool usage, we propose a task-frame, tool-centric action representation denoted as $T_{\text{tool}}^{\text{task}}$. This representation focuses on the tool being manipulated, independent of human or robot morphology or camera pose. A visual depiction of the coordinate systems is provided in Figure 3.

Using a 6D pose estimation model (e.g., FoundationPose [71]), we determine the tool's pose in the camera frame, $T_{\text{tool}}^{\text{camera}}$. To make the policy robust to camera and base movement, we transform this into the task frame:

$$T_{\text{tool}}^{\text{task}} = T_{\text{camera}}^{\text{task}} T_{\text{tool}}^{\text{camera}},$$

where $T_{\text{camera}}^{\text{task}}$ represents the transformation from the camera to the task frame.

E. Training and Deploying Robot Policies

We use diffusion policy [14] as our policy representation to predict $T_{\text{tool}}^{\text{task}}$, trained using the ACCESS system [7]. During deployment, for stationary robots, the task frame aligns with the base frame. For robots with moving base, base movement is compensated using $T_{\text{task}}^{\text{base}}$. The final end-effector pose in the robot base frame, used as the control command for the robot controller, is computed as:

$$T_{\text{eef}}^{\text{base}} = T_{\text{task}}^{\text{base}} T_{\text{tool}}^{\text{task}} T_{\text{eef}}^{\text{tool}},$$

where $T_{\text{eef}}^{\text{tool}}$ is the fixed transformation between the tool and the robot end-effector.

V. Policy Evaluations

Our experimental evaluations aim to assess the effectiveness of our framework for deploying robot policy learning across three key dimensions:

- **Capabilities:** What skills can our framework enable robots to acquire, and how robust are the policies under challenging conditions like a moving camera or base?
- **Effectiveness:** How well does our framework achieve its objectives outlined in Section IV? Can it support reliable and scalable policy learning for complex tasks while streamlining data collection for highly dynamic, contact-rich, or dexterous scenarios?
- **Policy Execution Efficiency:** How efficiently do the trained policies execute tasks? Does our framework enable smoother and more natural motion trajectories? Can it achieve faster and more fluid task completion compared to baseline methods?

To evaluate our framework for learning from direct human play, we developed a set of real-world robotic task domains using two embodiments: Kinova Gen3 and UR5e. These tasks are designed to test various aspects of policy capabilities, efficiency and effectiveness. Table I summarizes the key characteristics of these tasks. Visual input for the policies is provided by two RealSense D415 cameras. We describe the tasks in detail, highlighting their challenges and the specific capabilities we aim to test.

Nail Hammering: The task involves hammering a 3D-printed nail, requiring the robot to locate the nail, draw back the hammer, and strike the nail tip accurately. With a diameter of less than 15.5 mm, the nail tip demands high precision. Challenges include localizing the nail tip precisely and planning effective hammer trajectories. To evaluate generalization, the initial position of the nail is varied across different spatial configurations. We collected 180 seconds of data (40 human play episodes) from a single participant.

Meatball Scooping: In this task, the robot must use a spoon to scoop a meatball from a pan and transfer it to a bowl. This task is challenging due to the complex dynamics of the meatball, which can roll unpredictably within the pan. Additionally, the interaction between the spoon and the meatball requires careful control, as improper contact can cause the meatball to slip or escape the spoon. We randomize the initial position of the meatball within the pan to test its generalization capability. We collected 340 seconds of data (50 human play episodes) from a single participant.

Pan Flipping (Egg, Burger Bun, Meat Patty): The objective of this task is to use a pan to flip various objects, such as an egg, a burger bun, and a meat patty. The task is challenging due to its high-speed dynamics, requiring the robot to overcome gravity and accurately manage the interaction between the pan and the objects. Each object differs in weight, shape, and texture, adding further complexity. This task evaluates the policy's ability to handle fast, contact-rich interactions and adapt to diverse object types. To increase variability, the initial positions of the objects within the pan are randomized. Furthermore, the rapid and dynamic nature of the task makes it unsuitable for classical demonstration collection methods, highlighting the advantages of using bare-handed human play for data collection. We collected 50 seconds of data (38 human play episodes) from a single participant using three different pans and two object types.

Wine Balancing: In this task, the robot needs to use a hook to lift a wine bottle and carefully insert it into an unstable, zero-gravity wine rack. The task is challenging due to the precise control required to suspend the bottle in mid-air and counteract gravitational forces effectively. Any over-insertion or under-

TABLE I: **Benchmark Attributes of Real-World Tasks.** These benchmarks evaluate the precision, adaptability, and capability of our framework to address tasks requiring high precision, handling extreme dynamics, utilizing extrinsic dexterity, performing in contact-rich scenarios, and overcoming gravity.

Benchmark	High-Precision	Extreme Dynamics	Using Extrinsic Dexterity	Contact-Rich	Overcoming Gravity
Task 1: Nail Hammering	✓	–	–	–	–
Task 2: Meatball Scooping	✓	–	✓	✓	–
Task 3: Pan Flipping (Egg, Bun, Patty)	–	✓	✓	✓	✓
Task 4: Wine Balancing	✓	–	✓	✓	✓
Task 5: Soccer Ball Kicking	–	–	–	✓	–

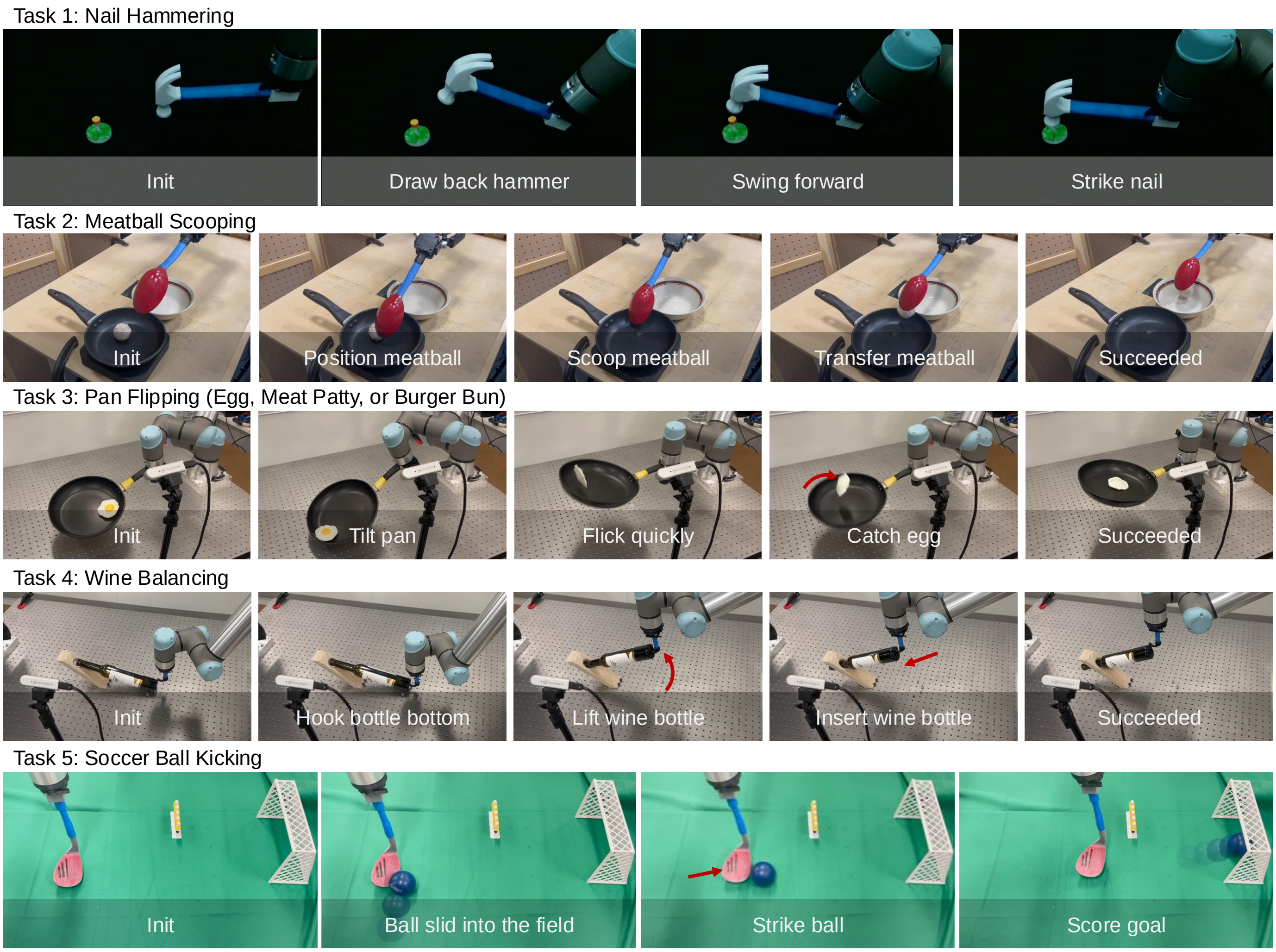


Fig. 4: **Policy Rollouts.** We evaluate diverse real-world tasks: nail hammering (precision in locating a nail tip), meatball scooping (slippery object, constrained environments), pan flipping (extremely dynamic, high-speed, contact-rich), wine balancing (precise control of unstable objects), and soccer ball kicking (dynamic object handling, goal-directed actions).

insertion will cause the bottle to lose balance. To constrain the horizontal movement of the rack, screws were added as obstacles to limit lateral motion. No additional variability was introduced. We collected 223 seconds of data (15 human play episodes) from a single participant.

Soccer Ball Kicking: In this task, the robot must use a golf club to kick a ball that slides into a field and direct it into the goal. To increase the challenge, a 3D-printed row of players serves as obstacles between the robot and the goal. The task is difficult because the robot must accurately intercept the moving ball, strike it with the correct force and direction, and ensure it avoids obstacles before reaching the goal. The position of the player obstacle varies. We collected 78 seconds of data (20 human play episodes) from a single participant.

Baselines: The primary focus of this work is to demonstrate the effectiveness and efficiency of learning directly from human play without relying on robot-generated data. We compare our approach against two baselines: a diffusion policy trained on robot demonstrations, and UMI [16], a hand-held gripper-based data collection method. The robot demonstration dataset is collected using either a SpaceMouse or Gello interface under the same data collection time. Additionally, we perform ablation studies to analyze the impact of key components, such as random cropping of images before policy training,

TABLE II: **Task Success Rates and Completion Times.** Success rates are the number of successful trials out of total episodes, and average completion times are based on successful trials. "DP" refers to the diffusion policy trained on teleoperation data. "Not Feasible" tasks denote cases where teleoperation failed due to extreme dynamics, precision, or reactivity demands. Our method consistently achieves higher success rates and shorter completion times.

Task	Method	Success Rate	Time (s)
Hammer Nailing	DP	0/13	-
	Ours	**13/13**	**11.0**
Meatball Scooping	DP	5/12	42.0
	Ours	**10/12**	**12.4**
Pan Flipping - Egg	DP	Not Feasible	-
	Ours	**12/12**	**1.5**
Pan Flipping - Burger Bun	DP	Not Feasible	-
	Ours	**9/12**	**1.9**
Pan Flipping - Meat Patty	DP	Not Feasible	-
	Ours	**10/12**	**2.3**
Wine Balancing	DP	Not Feasible	-
	Ours	**8/10**	**30.9**
Soccer Ball Kicking	DP	Not Feasible	-
	Ours	**6/10**	**2.0**

TABLE III: **Task success rates comparing our method with the hand-held gripper-based method on Nail Hammering.**

Method	Demo Duration & Count	Success Rate
UMI [16]	$\sim$180 seconds (25 demos)	0/13
UMI	$\sim$720 seconds (100 demos)	**13/13**
Ours	$\sim$180 seconds (40 demos)	**13/13**

novel view synthesis-based data augmentation, and the effects of embodiment segmentation. To further illustrate the advantages of our approach, we compare trajectory rollouts for a meatball-scooping episode, highlighting how our method is more sample-efficient and less prone to distribution shifts by eliminating excessive waypoints.

Evaluation Metrics: During testing, we introduce two types of variations: (1) randomizing the initial spatial configurations of objects in each task to assess policy generalization, and (2) varying camera positions to evaluate the robustness of policies to different viewpoints. All methods, including the baseline and ablation variants, are tested under the same conditions. Performance is evaluated using two metrics: success rate, which measures the proportion of successfully completed task trials and reflects policy effectiveness, and task completion time, which captures the average duration to complete tasks and reflects policy efficiency.

VI. Experiment Results

In our experiments, we demonstrate that our framework is both effective and efficient for training robots with advanced capabilities. Furthermore, leveraging human play data enables robots to perform smoother movements and acquire skills that are challenging or even impossible to achieve with robot-generated data.

A. Capabilities and Effectiveness

Table II presents the results of our real-world robot tasks, showing that our framework consistently outperforms baseline methods by achieving significantly higher success rates across all evaluated scenarios. We further compare our method with a stronger hand-held gripper baseline, UMI [16], as shown in Table III. In our default setup, SLAM-based mapping failed due to low environmental texture. To address this limitation, we added a textured background to support reliable mapping for UMI. For the nail hammering task, we evaluated UMI using 25 demonstrations (matching our collection time) and 100 demonstrations (to assess its ideal performance). UMI failed all 13 trials with 25 demonstrations but succeeded in all 13 trials with 100. UMI could not be applied to the wine balancing task due to contact-induced tool displacement, nor to the pan flipping task due to tool inertial slippage. In the soccer kicking task, large and fast motions made it nearly impossible to localize the demonstration trajectory within the initial map.

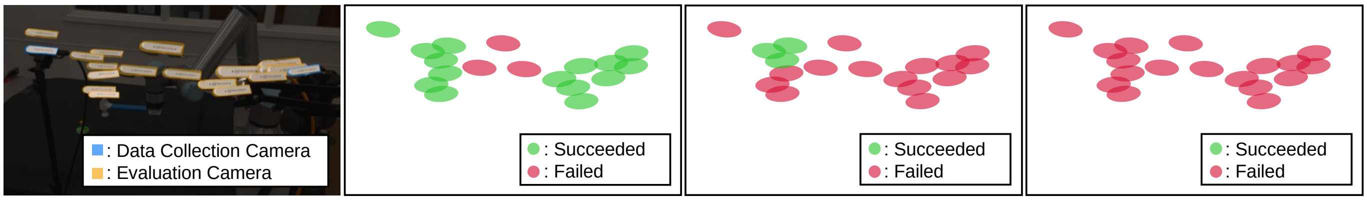
In contrast to the baselines, our method demonstrates reliable performance across all tasks. As illustrated in Figure 4, our method excels in real-world policy rollouts by accurately detecting spatial locations in tasks such as nail hammering and meatball scooping. For pan flipping tasks, it performs the high-speed motions required to effectively flip eggs, burger buns, and meat patties. Additionally, it demonstrates precise control in lifting and inserting a wine bottle into its stand and reacts swiftly and accurately to kick a soccer ball in the soccer ball kicking task. The superior performance of our framework is primarily due to its ability to collect a significantly larger volume of episodes within the same data collection timeframe, as detailed later in Section VII. This broader dataset covers a wider range of task variations, enabling more robust and adaptable policy training. Furthermore, our approach overcomes the inherent limitations of Gello and SpaceMouse, enabling the collection of demonstrations for scenarios that these tools cannot adequately handle due to their constraints.

B. Policy Execution Efficiency

Our framework demonstrates exceptional efficiency in task execution, achieving faster task completion times and producing smoother action motions compared to baseline methods, as shown in Table II. The efficiency is largely attributed to the nature of human play data, which captures the fluidity and speed of real-world human activities, resulting in smoother and more natural trajectories in the training dataset. In contrast, previous approaches relying on teleoperated data often suffer from significantly slower speeds and less natural motions, limiting their effectiveness in dynamic scenarios. By leveraging the realistic dynamics of human play, our framework not only accelerates task execution but also enhances motion quality, making it better suited for real-world applications.

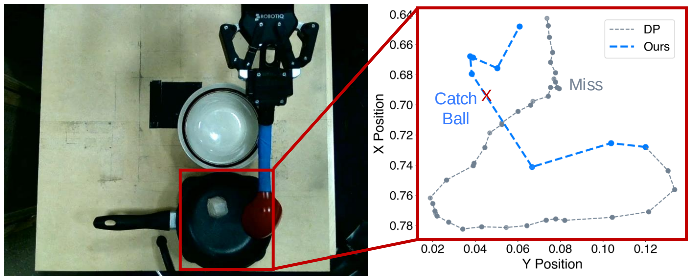
C. Benefits of Tool-Based Action Representation in Task Space

We observed that using the tool in the camera frame as the action representation allows the policy to perform comparably to prior works when the camera is stationary. However, the success rate drops to zero when the camera is moving. Tracking the camera pose in real time during motion



(a) Camera Poses (b) Ours w/ RC, w/ VA (c) Ours w/ RC, w/o VA (d) Ours w/o RC, w/o VA

Fig. 5: **Policy Testing Across Camera Poses in Nail Hammering.** (a) Camera poses for data collection and evaluation. (b-d) Performance ranges for methods trained with/without random cropping (RC) and view augmentation (VA).



(a) Initial Setup (b) EEF Trajectory Comparison

Fig. 6: **Policy Execution Trajectory Comparison.** (a) Initial setup for meatball scooping. (b) Comparison of end-effector XY trajectories from our framework and a policy trained on robot-collected data. Our method generates smoother, more natural, and efficient trajectories with fewer waypoints, enabling fluid and precise motion. In contrast, the robot-collected policy produces jerky movements with excessive waypoints, which hinder task performance.

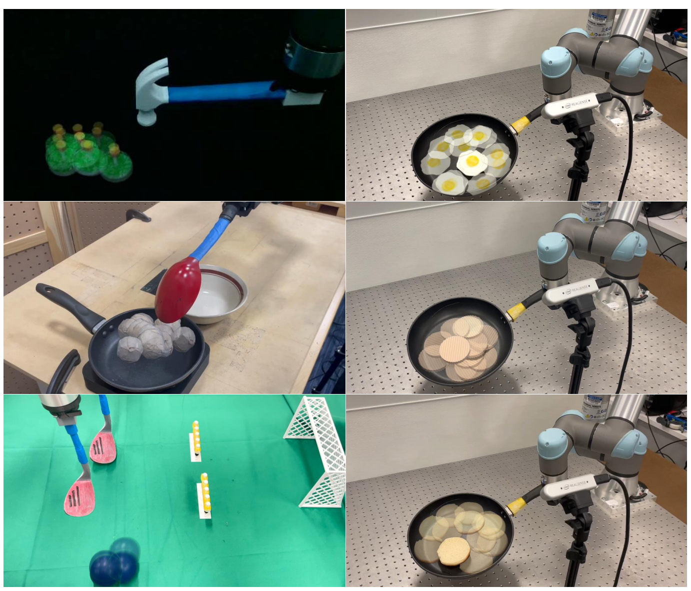


Fig. 7: **Initial States for All Evaluation Episodes.** All methods are evaluated using the same set of manually defined initial states, overlayed in the image. These states ensure diverse variations to test the policy's spatial generalization capabilities.

is non-trivial, leading to incorrect calculations of the end-effector position in the base frame. When the base moves, policies relying on the tool in the base frame fail entirely, achieving zero success. These failures occur because such policies assume a fixed relative transformation between the base and the workspace, an assumption invalidated by base motion. In contrast, representing actions using the tool in task space enables the policy to remain effective under base perturbations. Our approach demonstrates robust performance when camera poses vary, when the base moves, and even during simultaneous shaking of the camera and base.

D. Effects of Random Cropping and View Augmentation

Our experiments show that random cropping (RC) and view augmentation (VA) enhance policy robustness to diverse camera poses. Training with RC improves resilience to minor camera perturbations, such as small movements or shaking, while VA extends robustness by exposing the model to a broader range of camera configurations during training. We evaluated these techniques on the nail hammering task and present the results in Figure 5, comparing three models: one trained with both RC and VA, one with RC only, and one without either technique. The combined use of RC and VA significantly expands the working range of camera configurations, including those with substantial deviations from the original viewpoint. The robustness arises from the model's exposure to a diverse set of camera poses during training.

E. Policy Execution Trajectory Comparison

Our framework generates faster, smoother, and more natural trajectories compared to traditional approaches, as demonstrated by the end effector (EEF) XY trajectory for the meatball scooping task in Figure 6. Figure 6(a) depicts the initial setup of one trial, while Figure 6(b) contrasts the XY trajectory of the policy rollout using our framework with that of a policy trained on robot-collected data. Our framework surpasses the baseline approach in two critical aspects:

- **Improved Motion Fluidity:** Trajectories generated by our framework are significantly smoother and less jerky, enabling more fluid and precise task execution. In contrast, the robot-collected data results in jerky motions caused by excessive waypoints, which disrupt movement precision and degrade performance.
- **Efficient and Effective Learning:** The trajectories produced by our framework contain 10 times fewer waypoints than those from robot-collected data. The reduction simplifies the learning space, mitigates cumulative error growth, and minimizes the risk of distribution shifts commonly observed in behavior cloning. As a result, our policies exhibit significantly improved sample efficiency.

F. Generalization

Spatial Generalization: Our policy demonstrates strong generalization across various spatial configurations. Figure 7 illustrates the range of initial states tested for multiple tasks. In the

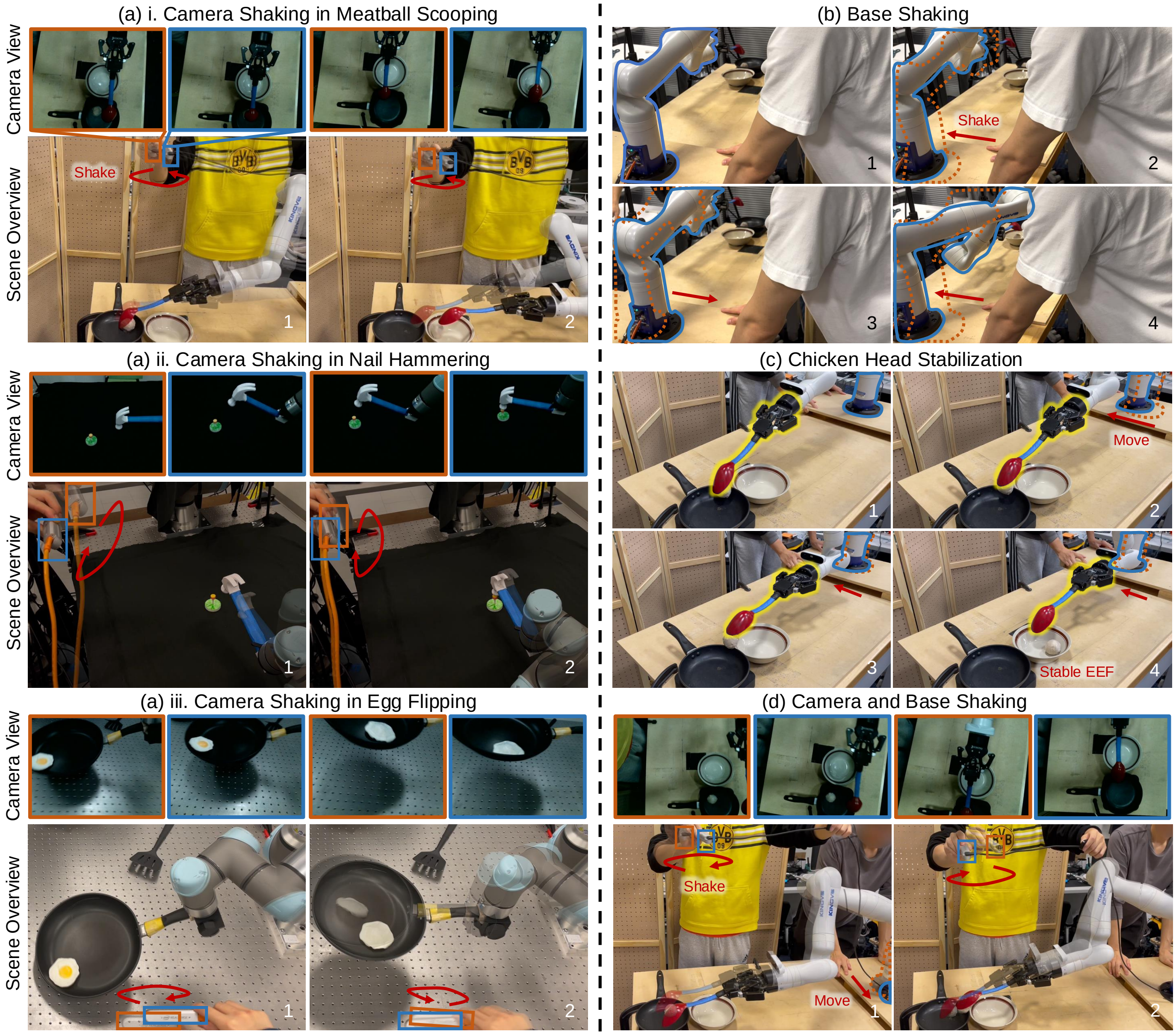


Fig. 8: **Robustness to Camera and Base Movement.** (a) Camera Pose Robustness: The policy demonstrated the ability to handle camera shaking across three tasks—meatball scooping, nail hammering, and pan flipping. The first row shows the camera view, while the second row provides a scene overview with the shaking motion. (b) Robot Base Robustness: The policy successfully compensated for base shaking, even when the shaking frequency exceeded the robot's control frequency. (c) Chicken Head Stabilization: At lower base movement frequencies, the end effector displayed a stabilization effect similar to a chicken's steady head. (d) Combined Robustness: The policy maintained task performance under simultaneous camera and base shaking.

hammer nailing task, nail positions vary. For meatball scooping, meatball positions are tested. In soccer ball kicking, we evaluate two goalkeeper configurations. For pan flipping, object poses vary to test all areas of the pan. As shown in Table II, our method achieves high success rates across these ranges.

Object Generalization: Our method generalizes effectively to different objects in pan flipping tasks, including a toy egg, a 3D-printed meat patty, and a real burger bun (Figure 7, second column). With only 13 demonstrations, the policy succeeds by leveraging a simple but effective strategy: tilting the pan to slide the object into a corner, then flicking the pan to propel and flip the object. Our manipulation approach for pan flipping enables robust generalization across diverse object types.

Tool Generalization: To assess the generalization ability of our policy across different tools, we conducted a pan-flipping experiment using a burger bun and five pans: large, medium, small, tiny, and square. For each pan, we collected 12 trials with varying initial configurations and reported the success rates (Fig. 10). The policy was trained on demonstrations using the large, medium, and square pans, and evaluated on all five. Results indicate that our method exhibits some generalization across both pan sizes and shapes (circular vs. square). High success rates were observed with the large and medium pans. However, performance declined on smaller pans, likely due to

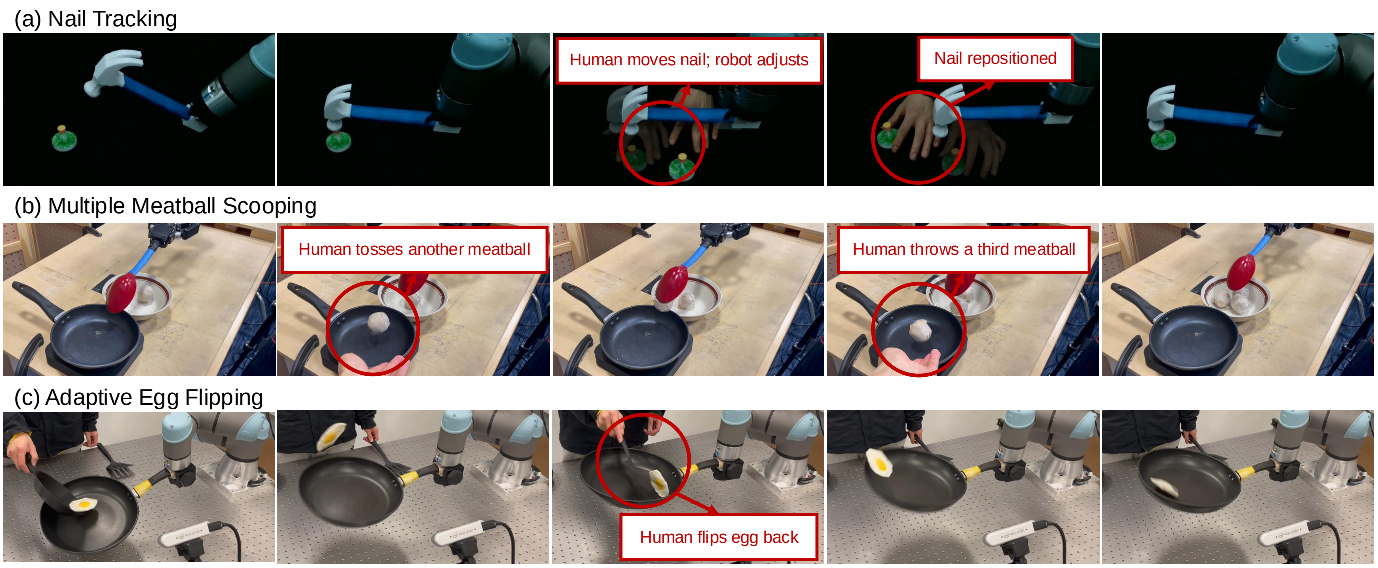


Fig. 9: **Human Perturbation Robustness.** This figure showcases the robot's ability to handle human-induced perturbations across three tasks: (1) In nail hammering, the robot successfully followed a manually moved nail; (2) In meatball scooping, it located and scooped meatballs even when additional ones were thrown into the pan mid-task; and (3) In egg flipping, the robot consistently flipped the egg back after human intervention repositioned it. These results highlight the policy's robustness to unpredictable human perturbations.

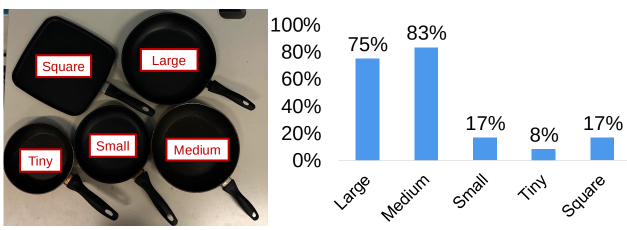


Fig. 10: **Tool Generalization** (a) The tested pans. (b) Success rate across 12 testing trials.

their limited surface area. The square pan also showed lower success rates, as its shallow edges allowed the bun to slide out during flipping.

G. Robustness

Camera Pose Robustness: We evaluated the policy's ability to handle camera pose variations by introducing camera shaking in three tasks: meatball scooping, nail hammering, and pan flipping (egg), as shown in Figure 8(a). The first row shows the camera view, while the second row shows the scene overview and the shaking motion. Despite the disturbances, the policy successfully completed all tasks. This robustness stems from random cropping during training, which allows the policy to adapt to partially cropped inputs and minor visual changes.

Robot Base Robustness: To evaluate robustness to base movement, we manually shook the robot base during task execution (Figure 8(b)). When the shaking frequency exceeded the control frequency, the end effector oscillated with the base. However, the task-space action design enabled the robot to compensate for these movements and maintain task success. Additionally, as shown in Figure 8(d), the policy remained effective even when both camera and base shaking were applied simultaneously.

Chicken Head Stabilization: At lower frequencies of base movement, where the shaking was slower than the robot's control frequency, the end effector exhibited a stabilization effect similar to a chicken's head remaining steady [75], as shown in Figure 8(c). The behavior highlights the robot's ability to maintain precise control during mild perturbations.

Human Perturbation Robustness: We further tested the policy's resilience to human-induced perturbations across three tasks, as shown in Figure 9. In the nail hammering task, the robot successfully tracked and followed a manually moved nail by a human. During meatball scooping, the robot located and scooped meatballs, even when additional meatballs were thrown into the pan mid-task. For egg flipping, the robot consistently flipped the egg back each time a human intervened and returned the egg to different initial positions. These results underscore the policy's robustness in maintaining reliable performance under real-time human interactions.

VII. Data Collection Efficiency and Affordability

We compare various data collection methods for robot imitation learning, focusing on throughput, reliability, cost, usability, and precision. Our evaluation includes teleoperation tools like Gello and Spacemouse for 6DOF (UR5e) and 7DOF (Kinova Gen3) robots, alongside methods such as Visual Imitation Made Easy, handheld grippers (e.g., UMI and LEGATO), and devices like VR (Meta Quest 2), AR (Apple Vision Pro), and Kinematic replicate (Gello).

A. Data Collection Efficiency

Our framework achieves significantly higher data collection throughput than traditional methods, enabling more demonstrations within the same timeframe. The improvement is driven by the natural and intuitive efficiency of human play, which ensures faster and more reliable task execution. Figure 11(a) highlights the superior manipulation capabilities of human hands, while Figure 12 quantifies the substantial

TABLE IV: **Comparison of Data Collection Methods.** This table compares various data collection methods for robotics. For cost, we calculate only the additional expenses required for data collection, excluding cameras, as they are considered a basic and commonly used sensor for robots rather than an additional purchase. Each method is assessed based on cost, ease of use, required expertise, precision, and maintenance effort. Our method stands out as cost-free, easy to use, highly precise, and requiring minimal maintenance.

Method	Cost	Ready-to-Use	Pre-Knowledge Required	Precise	Maintenance Expense
Visual Imitation Made Easy [78]	$340	No	Yes	No	Moderate
UMI [16]	$371	No	Yes	**Yes**	Moderate
LEGATO [57]	$1060	No	Yes	**Yes**	Moderate
Spacemouse [17]	$169	**Yes**	Yes	**Yes**	Low
VR (Meta Quest 2 [40])	$300	**Yes**	Yes	No	Moderate
AR (Apple Vision Pro [1])	$3499	**Yes**	Yes	**Yes**	High
Gello [74]	$272	No	Yes	No	Moderate
Ours	**$0**	**Yes**	**No**	**Yes**	**Minimal**

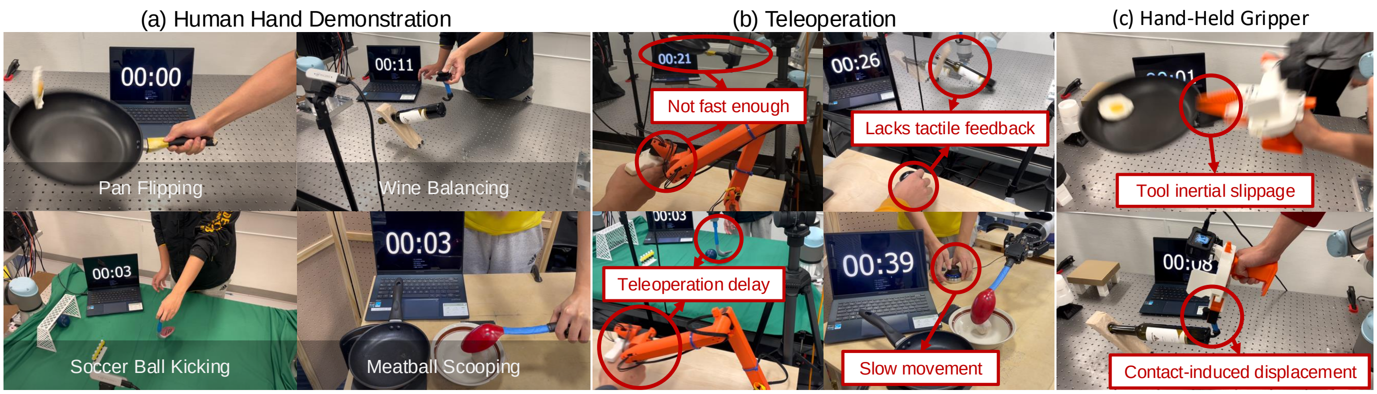


Fig. 11: **Data Collection Efficiency and Reliability.** (a) Human hands excel in manipulation tasks, leveraging natural and intuitive efficiency. (b) Failure cases for Gello and Spacemouse include insufficient speed, lack of tactile feedback during data collection, safety stops, collisions, teleoperation delays, and difficulty handling high-speed or complex tasks. (c) Failure cases for handheld grippers such as UMI [16], where issues arise from tool slippage due to inertia or displacement caused by contact forces.

time savings per episode. For nail hammering and meatball scooping, Gello and Spacemouse were used as teleoperation methods, respectively. Human hands reduced data collection time by 73% and 81% for nail hammering and meatball scooping, with consistently low variation in performance. In more complex tasks like pan flipping, wine balancing, and soccer ball kicking, teleoperation methods failed entirely due to limitations such as lack of tactile feedback, delays, and difficulty handling dynamic or precise actions. Our method further reduces data collection time by 41% compared to handheld grippers such as UMI [16] in nail hammering. UMI proved ineffective in wine balancing and pan flipping due to tool inertial slippage or contact-induced displacement, and failed in soccer kicking because of difficulty localizing large, fast motions. Moreover, it requires rich textures to build a pre-collection map, which our method does not. These results underscore the superior efficiency, robustness, and versatility of human play as a scalable solution for high-quality robot learning datasets.

B. Reliability

Figure 11(b) and Figure 11(c) illustrates typical failure cases with Gello, Spacemouse, and UMI [16], which frequently encounter issues such as safety stops or collisions during data collection. In contrast, our method ensures smooth, uninterrupted operation, avoiding these limitations. Traditional methods face significant challenges in high-speed or complex tasks. For example, Gello and Spacemouse struggle with replicating the extreme dynamics and precise motions required for flipping objects like eggs during pan flipping, often resulting in unsuccessful attempts. Similarly, teleoperation delays prevent timely strikes during soccer ball kicking, consistently leading to missed kicks and repeated failures. In tasks like wine balancing, the absence of tactile feedback impairs precision during the data collection, causing the wine bottle to tip over during data collection. Furthermore, in meatball scooping, the velocity vectors generated by Spacemouse input lead to jerky trajectories with redundant waypoints, significantly reducing efficiency. These challenges make effective training impractical with traditional methods. By leveraging human play, our framework not only addresses these limitations but also provides a reliable and scalable solution for dynamic and precision-demanding tasks.

C. Discussion of Data Collection Methods

Table IV compares various data collection methods based on cost, usability, expertise requirements, intuitiveness, and precision. Our method incurs no additional cost ($0), unlike hardware-dependent solutions like UMI and LEGATO, which demand significant investment. This affordability makes our approach accessible to users from diverse backgrounds without financial constraints. Unlike hardware-based systems such as UMI, LEGATO, Gello, and Spacemouse, which are prone to malfunctions and maintenance issues, our hardware-free framework ensures reliability and eliminates repair delays or

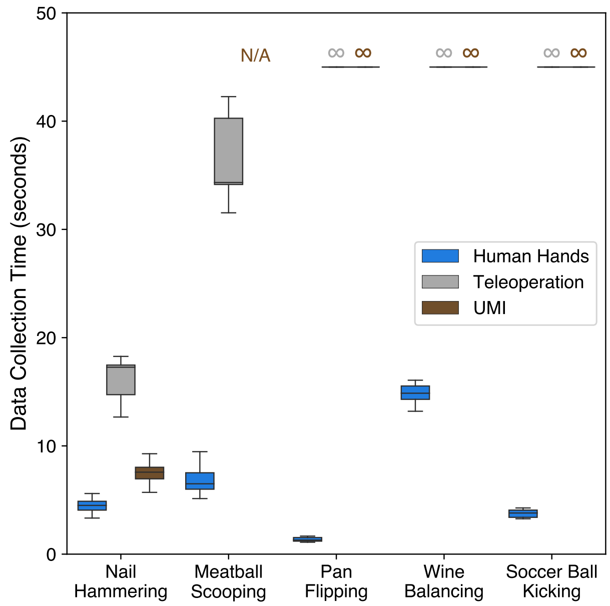


Fig. 12: **Quantitative Comparison of Data Collection Methods.** Human hands significantly reduce data collection time—by 73% for nail hammering and 81% for meatball scooping—while maintaining consistently low variation across trials. Teleoperation fails entirely in dynamic tasks and tasks requiring high precision. In nail hammering, human hands are 41% faster than UMI [16], which also struggles in dynamic scenarios and low-texture environments.

expenses. Additionally, it requires no supplementary 3D printing, in contrast to approaches like Visual Imitation Made Easy, UMI, and LEGATO. The simplicity of our design promotes inclusivity in collecting large-scale dataset for robot learning research. Our method also offers a more natural experience compared to tools like Spacemouse, while being far more cost-effective than VR and AR devices. Moreover, systems like Gello and Spacemouse lack the precision necessary for dynamic tasks, a limitation addressed by our approach. Overall, our method is a cost-effective, and accessible solution for data collection, overcoming key drawbacks of existing approaches while reducing complexity and maintenance needs.

VIII. Limitations and Future Work

Our framework has certain limitations. First, the perception pipeline relies on FoundationPose for extracting the tool's pose during manipulation. Errors in pose estimation may occasionally require data recollection, adding time and effort. Improving the reliability of the perception pipeline through more robust pose estimation algorithms or self-correction mechanisms is a promising direction for future work. Second, for novel view augmentation, significant noise and reduced realism are observed when augmented views deviate too far from the collected camera views, which can hinder policy performance. Future efforts could focus on leveraging advanced rendering techniques to enhance the realism of augmented views and improve policy generalization. Third, we assume the tool is rigidly attached to the robot's end effector; however, in real-world, contact-rich manipulation, minor shifts may occur, potentially affecting performance. Addressing this issue by incorporating tactile sensing could improve performance in contact-intensive tasks. Additionally, our method assumes a rigid tool and does not account for flexible or soft tools. Future work could explore using flexible representations for tool state estimation to better handle deformable tools in real-world manipulation scenarios.

IX. Conclusion

In this work, we presented a novel framework for human-to-robot imitation learning that leverages human play data to bridge the embodiment gap and enables robust policy training for diverse tool-use tasks. Unlike traditional data collection methods, which are often costly, hardware-dependent, and require technical expertise, our framework democratizes data collection by removing the need for specialized equipment or prior knowledge. Our approach makes data collection more accessible and scalable, empowering broader adoption in robotic learning. We validated our framework across a range of challenging tasks, including nail hammering, meatball scooping, pan flipping with various objects, wine bottle balancing, and soccer ball kicking. The results demonstrate the framework's superior performance, robustness to variations in camera poses and base movements, and adaptability to different embodiments, such as 6-DOF and 7-DOF robots. By enhancing accessibility, scalability, and reliability, our work lays a strong foundation for advancing robotic manipulation in complex, real-world scenarios.

Acknowledgments

We thank Professor Wenzhen Yuan, Xiaoyu Zhang, Yucheng Mo, Yilong Niu, Hyoungju Lim, and Amin Mirzaee for their support and assistance in reproducing the UMI. We thank Professor Saurabh Gupta, Professor Junyi Geng, Shuijing Liu, Lujie Yang, Neeloy Chakraborty, Hongkai Dai, Jacob Wagner, Shaoxiong Yao, Wei-Cheng Huang for their insightful feedback and suggestions. This work was supported by ZJU-UIUC Joint Research Center Project No. DREMES 202003, funded by Zhejiang University. This research used the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Delta is a joint effort of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications. Additionally, this work used NCSA Delta GPU at NCSA through allocation CIS240753 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

[1] Apple. Apple vision pro, 2024. URL https://www.apple.com/apple-vision-pro/.

[2] Maria Attarian, Muhammad Adil Asif, Jingzhou Liu, Ruthrash Hari, Animesh Garg, Igor Gilitschenski, and Jonathan Tompson. Geometry matching for multi-embodiment grasping, 2023.

[3] Arpit Bahety, Priyanka Mandikal, Ben Abbatematteo, and Roberto Martín-Martín. Screwmimic: Bimanual imitation from human videos with screw space projection. In *Robotics: Science and Systems (RSS), 2024*, 2024.

[4] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[5] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[6] A. Bicchi and V. Kumar. Robotic grasping and contact: a review. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, volume 1, pages 348–353 vol.1, 2000.

[7] Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, PEARC '23, page 173–176, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399852.

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[9] Anais Brygo, Ioannis Sarakoglou, Nadia Garcia-Hernandez, and Nikolaos Tsagarakis. Humanoid robot teleoperation with vibrotactile based balancing feedback. In *Haptics: Neuroscience, Devices, Modeling, and Applications: 9th International Conference, EuroHaptics 2014, Versailles, France, June 24-26, 2014, Proceedings, Part II 9*, pages 266–275. Springer, 2014.

[10] G. Carbone. *Grasping in Robotics*. Mechanisms and Machine Science. Springer London, 2012. ISBN 9781447146643.

[11] Boyuan Chen, Pieter Abbeel, and Deepak Pathak. Unsupervised learning of visual 3d keypoints for control. In *ICML*, 2021.

[12] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[13] Tao Chen, Adithyavairavan Murali, and Abhinav Gupta. Hardware conditioned policies for multi-robot transfer learning, 2019.

[14] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[15] 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*, 2024.

[16] 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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[17] 3D Connextion. Spacemouse, 2023. URL https://3dconnexion.com/us/spacemouse/.

[18] Coline Devin, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, and Sergey Levine. Learning modular neural network policies for multi-task and multi-robot transfer. In *International Conference on Robotics and Automation (ICRA)*.

[19] Kiran Doshi, Yijiang Huang, and Stelian Coros. On hand-held grippers and the morphological gap in human manipulation demonstration. *arXiv preprint arXiv:2311.01832*, 2023.

[20] 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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[21] Bin Fang, Shidong Jia, Di Guo, Muhua Xu, Shuhuan Wen, and Fuchun Sun. Survey of imitation learning for robotic manipulation. *International Journal of Intelligent Robotics and Applications*, 3(4):362–369, Dec 2019. ISSN 2366-598X.

[22] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[23] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[24] 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 *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[25] Edward S. Hu, Kun Huang, Oleh Rybkin, and Dinesh Jayaraman. Know thyself: Transferable visual control policies through robot-awareness. In *International Conference on Learning Representations (ICLR)*, 2022.

[26] Jiang Hua, Liangcai Zeng, Gongfa Li, and Zhaojie Ju. Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors*, 21(4), 2021. ISSN 1424-8220.

[27] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2), April 2017. ISSN 0360-0300.

[28] Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

[29] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, volume 164, pages 991–1002. PMLR, 2022.

[30] P. Jonsson. Ericsson mobility report november 2024. In *Ericsson Mobility Report*. Ericsson, 2024.

[31] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *ArXiv*, abs/2410.24221, 2024.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[33] J. Kofman, Xianghai Wu, T.J. Luu, and S. Verma. Teleoperation of a robot manipulator using a vision-based human-robot interface. *IEEE Transactions on Industrial Electronics*, 52(5):1206–1219, 2005.

[34] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.

[35] Andrew Lobbezoo, Yanjun Qian, and Hyock-Ju Kwon. Reinforcement learning for pick and place operations in robotics: A survey. *Robotics*, 10(3), 2021. ISSN 2218-6581.

[36] T. Lozano-Perez, J.L. Jones, E. Mazer, and P.A. O'Donnell. Task-level planning of pick-and-place robot motions. *Computer*, 22(3):21–29, 1989.

[37] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. PMLR, 2020.

[38] 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. In *The Eleventh International Conference on Learning Representations*, 2023.

[39] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning (CoRL)*, volume 87, pages 879–893. PMLR, 2018.

[40] Meta. Meta quest, 2023. URL https://www.meta.com/quest.

[41] François Osiurak and Dietmar Heinke. Looking for intoolligence: A unified framework for the cognitive study of human tool use and technology. *American Psychologist*, 73(2):169–185, 2018.

[42] Chuer Pan, Brian Okorn, Harry Zhang, Ben Eisner, and David Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 1783–1792. PMLR, 2023.

[43] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[44] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3810, 2018.

[45] Luka Peternel and Jan Babič. Learning of compliant human–robot interaction using full-body haptic interface. *Advanced Robotics*, 27(13):1003–1012, 2013.

[46] Domenico Prattichizzo and Jeffrey C. Trinkle. *Grasping*, pages 955–988. Springer International Publishing, Cham, 2016. ISBN 978-3-319-32552-1.

[47] Pragathi Praveena, Guru Subramani, Bilge Mutlu, and Michael Gleicher. Characterizing input methods for human-to-robot demonstrations. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 344–353. IEEE, 2019.

[48] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

[49] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[50] Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task-agnostic offline reinforcement learning. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 1838–1849. PMLR, 2023.

[51] Fereshteh Sadeghi, Alexander Toshev, Eric Jang, and Sergey Levine. Sim2real viewpoint invariant visual servoing by recurrent control. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4691–4699, 2018.

[52] Gautam Salhotra, xI Chun Arthur Liu, and Gaurav Sukhatme. Bridging action space mismatch in learning from demonstrations. *arXiv preprint arXiv:2304.03833*, 2023.

[53] Felipe Sanches, Geng Gao, Nathan Elangovan, Ricardo V Godoy, Jayden Chapman, Ke Wang, Patrick Jarvis, and

Minas Liarokapis. Scalable. intuitive human to robot skill transfer with wearable human machine interfaces: On complex, dexterous tasks. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6318–6325. IEEE, 2023.

[54] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. *CVPR, 2024*, 2023.

[55] Karl Schmeckpeper, Annie Xie, Oleh Rybkin, Stephen Tian, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*, pages 708–725. Springer, 2020.

[56] Mingyo Seo, Steve Han, Kyutae Sim, Seung Hyeon Bang, Carlos Gonzalez, Luis Sentis, and Yuke Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2023.

[57] Mingyo Seo, H. Andy Park, Shenli Yuan, Yuke Zhu, , and Luis Sentis. Legato: Cross-embodiment visual imitation using a grasping tool, 2024.

[58] Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 486–487, 2017.

[59] Carmelo Sferrazza, Dun-Ming Huang, Fangchen Liu, Jongmin Lee, and Pieter Abbeel. Body transformer: Leveraging robot embodiment for policy learning. In *8th Annual Conference on Robot Learning*, 2024.

[60] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[61] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[62] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 654–665. PMLR, 2023.

[63] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[64] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022.

[65] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *Robotics and Automation Letters*, 2020.

[66] Alexander Toedtheide, Xiao Chen, Hamid Sadeghian, Abdeldjallil Naceri, and Sami Haddadin. A force-sensitive exoskeleton for teleoperation: An application in elderly care robotics. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12624–12630. IEEE, 2023.

[67] Krist Vaesen. The cognitive bases of human tool use. *Behavioral and Brain Sciences*, 35(4):203–218, 2012.

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

[69] Tingwu Wang, Renjie Liao, and Sanja Fidler Jimmy Ba. Nervenet: Learning structured policy with graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[70] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[71] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024.

[72] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[73] Josiah Wong, Albert Tung, Andrey Kurenkov, Ajay Mandlekar, Li Fei-Fei, and Silvio Savarese. Error-aware imitation learning from teleoperation data for mobile manipulation. In *Conference on Robot Learning*, 2021.

[74] Philipp Wu, Fred Shentu, Xingyu Lin, and Pieter Abbeel. GELLO: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023*, 2023.

[75] Shuyan Xia, Yusen Li, Guilin Wen, Daolin Xu, Kai Wang, and Haicheng Zhang. Natural mechanism of superexcellent vibration isolation of the chicken neck. *Journal of Sound and Vibration*, 594:118649, 2025. ISSN 0022-460X.

[76] 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, 2024.

[77] Jonathan Heewon Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. In *Annual Conference on Robot Learning (CoRL)*, 2023.

[78] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy, 2020.

[79] Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. In *8th Annual Conference on Robot Learning*, 2024.

[80] Wenhao Yu, Jie Tan, C. Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. In *Robotics: Science and Systems (RSS)*, 2017.

[81] Wenhao Yu, Visak CV Kumar, Greg Turk, and C. Karen Liu. Sim-to-real transfer for biped locomotion. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3503–3510, 2019.

[82] Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. In *8th Annual Conference on Robot Learning*, 2024.

[83] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[84] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635. IEEE, 2018.

[85] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[86] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744, 2020.

[87] Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. NeRF in the Palm of Your Hand: Corrective Augmentation for Robotics via Novel-View Synthesis . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17907–17917, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.

[88] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[89] Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. *ArXiv*, abs/2310.14386, 2023.

Appendix A
Appendix

A. Implementation Details

1) Hardware Design: We designed two fast tool changers compatible with robots using the ISO 9409-1-50-4-M6 flange, as shown in Figure 13. The left design utilizes a screw mechanism to accommodate general tools, while the right design employs clips for tools with specific mounting shapes.

Fig. 13: **Fast Tool Changer.** Two designs are shown: the left accommodates general tools with a screw mechanism, and the right clips onto tools with specific mounting shapes.

2) Tool Pose Estimation: We use Polycam to scan the tool and obtain its mesh. The mesh is later feed into FoundationPose [71] for 6D pose estimation.

B. Additional Results

1) Effects of Embodiment Segmentation: Embodiment Segmentation masks the agent's embodiments during data collection and policy deployment, ensuring visually consistent scenes and reducing the training-deployment visual gap. Embodiment Segmentation significantly improves policy performance, as shown in Figure 14. Figure 14(a) highlights failure cases without segmentation. In the wine balancing task, the robot strikes the table, triggering safety stops due to improper bottle handling. In the soccer ball kicking task, the robot's actions are inconsistent, shorter, and less precise than during training. Quantitative results in Figure 14(b) further underscore segmentation's impact. Across 10 trials, segmentation enabled 8 successes in the wine balancing task, while the model without it achieved none. Similarly, in the soccer ball kicking task, segmentation resulted in 6 successes, compared to 2 without it. By aligning training and testing visual distributions, Embodiment Segmentation ensures consistent and reliable robot performance during the training and deployment.

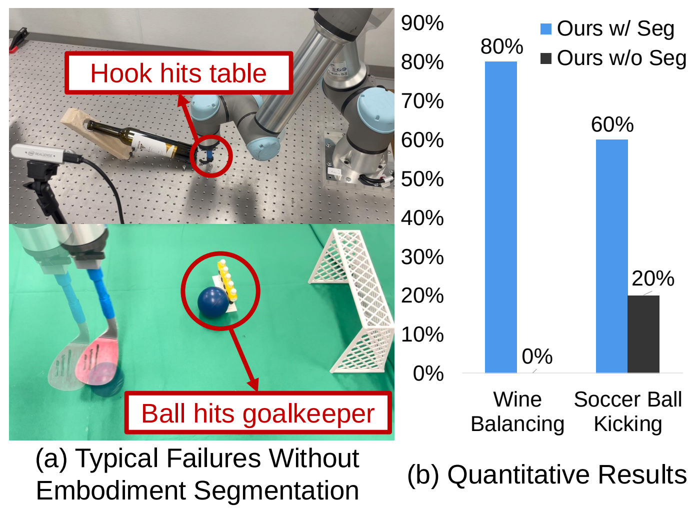


Fig. 14: **Effects of Embodiment Segmentation.** (a) Failure cases without segmentation: In the wine balancing task, the robot strikes the table, triggering safety stops. In the soccer ball kicking task, it performs shorter, less precise actions. (b) Quantitative results: Segmentation improved success rates in wine balancing (8 vs. 0) and soccer ball kicking (6 vs. 2) by reducing the visual gap between training and deployment.