Tool-as-Interface: Learning Robot Policies from Observing Human Tool Use

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

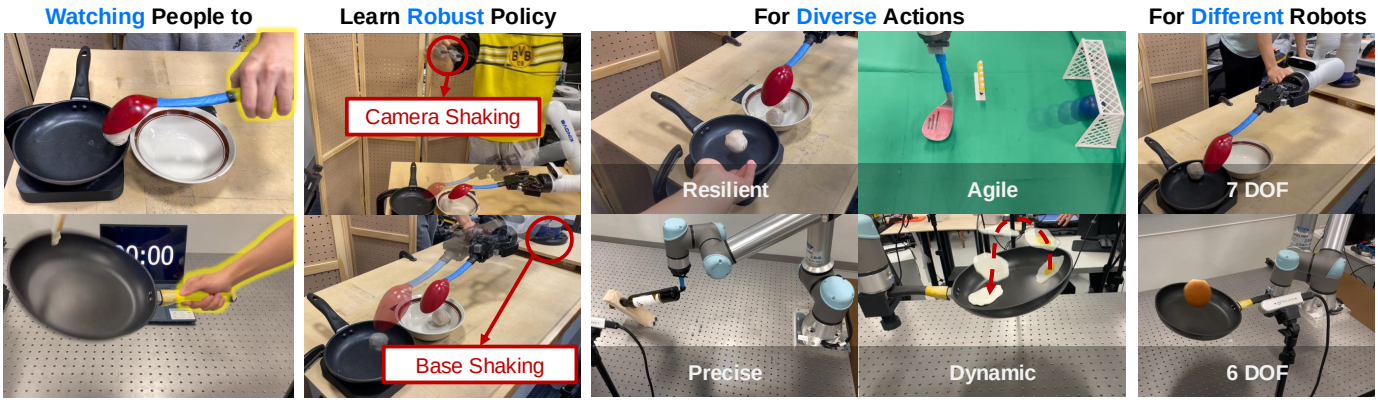


Fig. 1: **Tool-as-Interface.** We propose a scalable data collection and policy learning framework designed to transfer diverse, intuitive, and natural human play into effective visuomotor policies. The framework enables robots to learn robust policies that can operate effectively under challenging conditions,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

***Abstract*—Tool use is essential for enabling robots to perform complex real-world tasks, but learning such skills requires extensive datasets. While teleoperation is widely used, it is slow, delay-sensitive, and poorly suited for dynamic tasks. In contrast, human videos provide a natural way for data collection without specialized hardware, though they pose challenges on robot learning due to viewpoint variations and embodiment gaps. To address these challenges, we propose a framework that transfers tool-use knowledge from humans to robots. To improve the policy's robustness to viewpoint variations, we use two RGB cameras to reconstruct 3D scenes and apply Gaussian splatting for novel view synthesis. We reduce the embodiment gap using segmented observations and tool-centric, task-space actions to achieve embodiment-invariant visuomotor policy learning. Our method achieves a 71% improvement in task success and a 77% reduction in data collection time compared to diffusion policies trained on teleoperation with equivalent time budgets. Our method also reduces data collection time by 41% compared with the state-of-the-art data collection interface.**

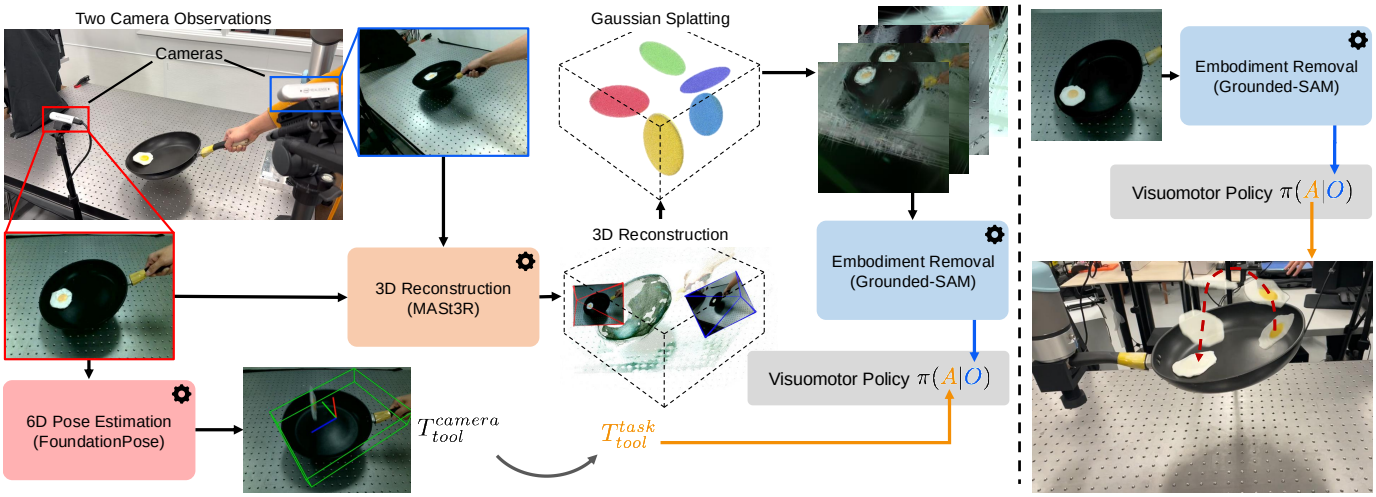
I. Introduction

Tool use enables humans to perform complex tasks by extending their physical capabilities. In contrast, robotic systems remain largely limited to grasping and pick-and-place operations [24, 18, 3, 20, 5, 16, 15]. To enable richer manipulation skills, robots must learn to use diverse tools in dynamic environments. This work focuses on the efficient training of robot policies for tool use, with an emphasis on scalable and low-cost data collection.

Imitation learning (IL) provides a promising approach for acquiring tool-use skills directly from human demonstrations [10, 11, 12]. Prior work has leveraged teleoperation platforms [25, 4, 12] and hand-held grippers [22, 8] to provide precise supervision. However, these systems often require expensive hardware, 3D-printed tools, or expert calibration, limiting their use beyond controlled environments. Although effective, these methods are difficult to scale.

Natural human manipulation videos—capturing everyday tool use without specialized equipment—offer a scalable and intuitive alternative for data collection. These demonstrations require no robotic infrastructure or technical preparation, yet remain underutilized in IL due to embodiment mismatch and the different perspective of single-view recordings [23, 2, 19]. We introduce a new framework that leverages two-view human manipulation videos to train robot policies. Using 3D scene reconstruction and novel view synthesis, the framework enables viewpoint-invariant learning. Embodiment-specific cues are filtered using segmentation, and a task-space, tool-centric action representation supports robustness to robot base variation (Figure 1).

Our contributions are as follows: (1) we introduce a framework for scalable, intuitive, and cost-effective data collection for robot tool-use learning, using two-view human manipulation videos without requiring teleoperation or specialized hardware; (2) we demonstrate strong generalization across diverse real-world tool-use tasks (e.g., nail hammering, meatball scooping, pan flipping, wine bottle balancing, and soccer ball kicking) achieving a 71% higher success rate and 77% reduction in data collection time compared to diffusion policies trained on SpaceMouse [9] or Gello [27], and a 41% improvement over handheld grippers like UMI [8]; and (3) we



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

Fig. 2: **Policy Design.** Human manipulation data was collected using two cameras and processed through the foundation model MASt3R [14] to generate 3D reconstructions. Using 3D Gaussian splatting, we sampled novel views to augment the dataset. Human hands were segmented to create embodiment-agnostic observations as policy inputs. For action labeling, FoundationPose [26] estimated the tool's pose in the camera frame, $T_{\text{tool}}^{\text{camera}}$, which was transformed into task space, $T_{\text{tool}}^{\text{task}}$. A diffusion model was then trained as the visuomotor policy.

provide a detailed robustness analysis, evaluating performance under changes in viewpoint, robot base configuration, and human motion, along with ablations on segmentation, novel view synthesis, and random cropping.

II. Tool-as-Interface Framework

Problem Statement: We formulate robotic manipulation as a Markov Decision Process (MDP), where the goal is to learn a policy $\pi : \mathcal{O}^r \to \mathcal{A}$ that enables a robot to perform a given task. The robot's observation space $\mathcal{O}^r$ consists of a single-view RGB image $I^r \in \mathcal{I}^r$ and proprioceptive data $x^r \in SE(3)$, where each I^r is a tensor in $\mathbb{R}^{128\times128\times3}$. We train the policy using an imitation dataset of N human demonstrations, $D = (\mathcal{O}_0^h, \mathcal{O}_1^h, \dots)_{n=1}^N$, where each $\mathcal{O}^h = \{I_{v1}^h, I_{v2}^h\}$ contains two RGB images captured from different viewpoints and each $I_{vi}^h \in \mathcal{I}^h$ is a tensor in $\mathbb{R}^{480\times640\times3}$. We preprocess the dataset to infer actions using a 6D pose estimation and tracking model, resulting in $D = \{(\mathcal{O}_0^h, a_0, \mathcal{O}_1^h, a_1, \dots)\}_{n=1}^N$, where each action $a \in SE(3)$. To bridge the embodiment gap between humans and robots, we assume the tool is rigidly attached to both the human hand (implicitly) and the robot end-effector (explicitly), with a fixed transformation estimated prior to deployment. Under this setup, the robot can reproduce human-demonstrated tool trajectories, enabling policy transfer across embodiments while preserving task-relevant behaviors (Figure 2).

Tool-Centric Demonstrations for Robot Manipulation: We leverage the fact that both humans and robots can operate the same physical tools to facilitate policy learning. Tools serve as a shared interface for interacting with objects, enabling the direct transfer of human demonstrations with minimal embodiment-specific adaptation. Unlike prior work focused on grasping or pick-and-place tasks [8, 22, 29], our approach enables robots to perform complex interactions using everyday tools. Our formulation abstracts actions to the tool pose, reducing morphological dependence and promoting policy generalization across embodiments. It also simplifies data collection by eliminating the need for robot-specific demonstrations. Humans can naturally manipulate tools by hand without extra instrumentation. For deployment, robots either rigidly grasp the tool, as demonstrated with a Kinova Gen3 arm, or attach it using a custom fast tool changer described in Appendix B2 and shown in Figure 6, compatible with ISO 9409-1-50-4-M6 flanges.

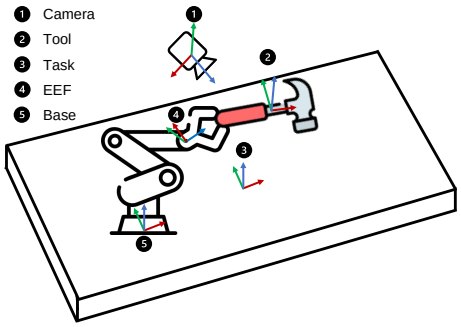


Fig. 3: **Coordinate System Diagram.** The diagram shows the Camera, Tool, Task space, End-Effector (EEF), and Base frames. The action is represented as $T_{\text{tool}}^{\text{task}}$.

Perception Alignment Across Embodiments: To enable cross-embodiment policy transfer, we align human and robot observations within a shared visual space $\mathcal{I}^s$ by applying a feature extraction function $g : \mathcal{I}^h \cup \mathcal{I}^r \to \mathcal{I}^s$. We instantiate g with Grounded-SAM [21], using prompts such as "human hand" and "robot arm" to mask out embodiment-specific regions—human hands during training and robotic arms during deployment. Masking out these regions ensures that only task-relevant visual information (e.g., tools and objects) remains visible in both phases. By minimizing visual discrepancies between human and robot data, the feature extraction process reduces embodiment-specific bias and improves generalization across embodiments.

3D-Aware View Augmentation: We use cameras for data collection due to their widespread availability—over 7.14 billion smartphones are equipped with them [13]. However,

single-camera setups suffer from scale ambiguity and limited 3D perception and are sensitive to viewpoint changes.

3D RECONSTRUCTION: To address this, we use MASt3R [14], an image-matching model that reconstructs accurate 3D environments from just two RGB images—eliminating the need for depth sensors, which are less common and more power-hungry. Two cameras suffice to avoid scale ambiguity inherent in monocular settings. MASt3R produces high-quality point clouds without requiring known camera extrinsics or intrinsics by globally aligning multi-view features.

VIEW SYNTHESIS AND AUGMENTATION: 3D Gaussian splatting synthesizes novel viewpoints from the reconstructed scene, allowing the robot to observe interactions from multiple angles—even when only two views are available. The resulting perspectives augment the training data, increasing visual diversity and improving policy learning. To further enhance robustness and generalization, random cropping is applied, following diffusion policy [6, 7].

Tool-Centric Action Representation and Policy Deployment: To support general tool usage, we propose a task-frame, tool-centric action representation denoted as $T_{\text{tool}}^{\text{task}}$, which describes the tool's motion independently of human or robot morphology, camera pose, or base configuration. This invariant formulation enables robust policy transfer across different embodiments and viewpoints. As shown in Figure 3, the tool's pose is first estimated in the camera frame using a 6D pose estimation model (e.g., FoundationPose [26]) as $T_{\text{tool}}^{\text{camera}}$, and then transformed into the task frame:

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

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

A diffusion policy [6] maps a single-view RGB image to a predicted SE(3) action $T_{\text{tool}}^{\text{task}}$. At deployment, the robot command is computed by converting the prediction to the end-effector frame. For stationary robots, the task frame aligns with the base frame; for mobile platforms, base movement is compensated using $T_{\text{task}}^{\text{base}}$. The resulting end-effector pose is given by:

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

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

III. POLICY EVALUATIONS

Our evaluations aim to assess our framework across three dimensions: **reliability** (how consistently and successfully the learned policies perform), **execution efficiency** (how smooth and natural the resulting behaviors are), and **versatility** (how well the framework adapts to diverse tasks and generalizes across conditions).

Experimental Tasks Overview: We evaluate five real-world robotic tasks on Kinova Gen3 and UR5e robots, involving precision manipulation, dynamic object handling, and dexterous tool use. Policies use RGB inputs from RealSense D415 cameras and handle variations in object positions and camera poses. Tasks include: (1) Nail Hammering – Precise striking of a small target, (2) Meatball Scooping – Contact-sensitive rolling object manipulation, (3) Pan Flipping – Fast, dynamic flipping with varied objects, (4) Wine Balancing – Gravity-aware placement into an unstable rack, and (5) Soccer Ball Kicking – Dynamic interception and obstacle avoidance. Full details in Appendix B1.

TABLE I: **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**

Baselines: We evaluate the effectiveness and efficiency of learning directly from human manipulation videos without relying on robot-generated data. We benchmark against a diffusion policy trained on robot demonstrations and UMI [8], a hand-held gripper method. Robot demonstrations are collected using SpaceMouse or Gello [27] under identical time budgets. Additionally, we conduct ablations to assess random cropping before policy training, novel view synthesis data augmentation, and 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.

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

Method	Demo Duration & Count	Success Rate
UMI [8]	∼180 seconds (25 demos)	0/13
UMI	∼720 seconds (100 demos)	**13/13**
Ours	∼180 seconds (40 demos)	**13/13**

IV. Experiment Results

Capabilities and Effectiveness: Table I summarizes our real-world results, showing that our framework consistently achieves higher success rates across all tasks compared to baselines. We also compare against the stronger hand-held gripper baseline UMI [8] (Table II). In our default setup, SLAM-based mapping failed due to low environmental texture, so we added a textured background to support reliable mapping for UMI. For the nail hammering task, we evaluated UMI with 25 demonstrations (matching our collection time) and 100 demonstrations (to assess ideal performance). UMI fails all 13 trials with 25 demonstrations but succeeded with 100. It was also inapplicable to wine balancing and pan flipping due to contact and inertial challenges, and struggled in soccer kicking due to localization failures. In contrast, our method demonstrates reliable performance across all tasks: accurately detecting spatial locations (nail hammering, meatball scooping), performing high-speed motions (pan flipping), precisely inserting wine bottles, and swiftly reacting in soccer kicking. This strong performance is enabled by collecting significantly larger and more diverse episodes within the same data collection timeframe, enabling robust policy training. Our approach overcomes limitations of teleoperation tools like Gello and SpaceMouse, enabling data collection for scenarios they struggle to handle. Qualitative policy rollouts are shown in Figure 5 in Appendix.

Generalization:

Spatial Generalization: We evaluated spatial generalization by varying initial conditions across tasks: nail positions for hammering, meatball locations for scooping, goalkeeper setups for soccer ball kicking, and object poses across the pan for flipping (illustrated in Figure 7 in Appendix).

Object Generalization: Our method generalizes effectively to different objects in the pan-flipping task, including the egg and burger bun seen during training, and a 3D-printed meat patty (illustrated in Figure 7, second column, in Appendix). The policy learns to tilt the pan to slide the object into a corner, then flick it to achieve a successful flip, enabling robust generalization across object types.

Tool Generalization: We evaluated tool generalization by testing the policy with five different pans: large, medium, small, tiny, and square. The policy was trained using demonstrations with the large, medium, and square pans and evaluated on all five, with 12 trials per pan under varying initial configurations (illustrated in Figure 4 in Appendix). It achieved high success rates on the trained pans (large and medium). Performance declined on smaller pans due to limited surface area, and on the square pan due to shallow edges causing the bun to slide out during flipping.

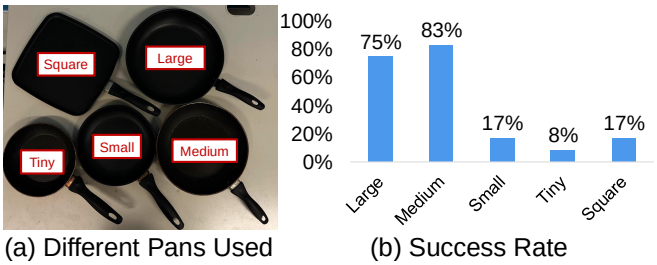


Fig. 4: **Tool Generalization.** (a) The tested pans. (b) Success rate across 12 testing trials.

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 (Figure 12(a)). The first row shows the camera view, and the second shows the scene overview and shaking motion. Despite disturbances, the policy consistently completed all tasks, enabled by random cropping during training, improving adaptation to partial views and minor visual changes.

Robot Base Robustness: To assess robustness to base movement, we manually shook the robot base during execution (Figure 12(b)). When the shaking frequency exceeded the control frequency, the end effector oscillated with the base; however, the task-space action design enabled compensation and successful task completion. As shown in Figure 12(d), the policy also maintained effectiveness under simultaneous camera and base shaking.

Chicken Head Stabilization: At lower shaking frequencies, where the perturbation was slower than the robot's control loop, the end effector exhibited a stabilization behavior similar to a chicken's head [28] (Figure 12(c)), maintaining steady control during mild base movements.

Human Perturbation Robustness: We evaluated resilience to human interventions (Figure 10). The robot tracked moving nails, adapted to new meatballs thrown in mid-task, and re-flipped repositioned eggs, demonstrating robustness to real-time disturbances.

V. Conclusion

In this work, we presented a framework for human-to-robot imitation learning that leverages human manipulation video to bridge the embodiment gap and enable robust policy training for diverse tool-use tasks. Unlike traditional data collection methods, which are often costly and hardware-dependent, our approach democratizes data collection by eliminating the need for specialized equipment or technical expertise, making large-scale robot learning more accessible and scalable. We validated the framework across challenging tasks, including nail hammering, meatball scooping, pan flipping, wine bottle balancing, and soccer ball kicking, demonstrating superior performance, robustness to variations in camera poses and base movements, and adaptability across 6-DOF and 7-DOF robots. By improving 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, 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] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 3D Connextion. Spacemouse, 2023. URL https://3dconnexion.com/us/spacemouse/.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

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

[25] 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.

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

[27] 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.

[28] 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.

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

APPENDIX

APPENDIX

CONTENTS

I Introduction 1

II Tool-as-Interface Framework 2

III Policy Evaluations 3

IV Experiment Results 4

V Conclusion 4

Appendix 7

A Design Choice 7

A1 Key Capabilities and Practical Benefits 7

B Detailed Experiment Setup 7

B1 Task Descriptions 7

B2 Implementation Details . . . 9

C Additional Experimental Results 9

D Detailed Analysis on Data Collection Efficiency and Affordability 10

D1 Data Collection Efficiency . 10

D2 Reliability 11

D3 Discussion of Data Collection Methods 12

A. Design Choice

1) Key Capabilities and Practical Benefits: Our framework enables the direct transfer of human manipulation data into deployable robot policies. It is designed to fulfill the following key objectives:

- **Support for Dynamic and High-Precision Tasks:** Human manipulation, 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 [13]. 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.

B. Detailed Experiment Setup

1) Task Descriptions: **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 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,

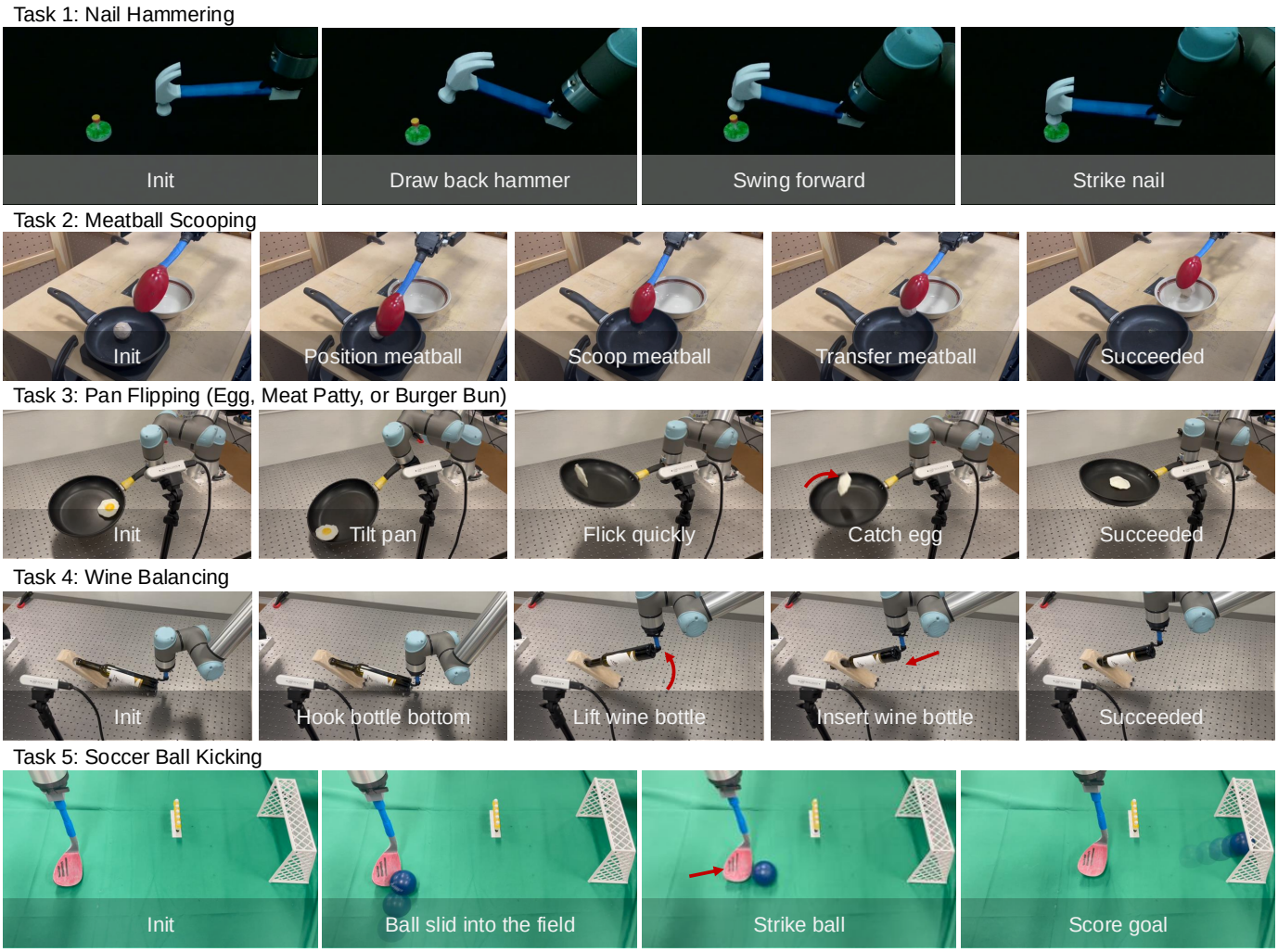


Fig. 5: **Policy Rollouts.** We evaluate diverse real-world tasks: nail hammering (precision in locating a nail tip), meatball scooping (slippery object, constrained environments), pan flipping (extremely dynamic, high-speed, contact-rich), wine balancing (precise control of unstable objects), and soccer ball kicking (dynamic object handling, goal-directed actions).

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 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 videos for data collection. We collected 50 seconds of data (38 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-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 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 episodes) from a single participant.

TABLE III: **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. 6: **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.

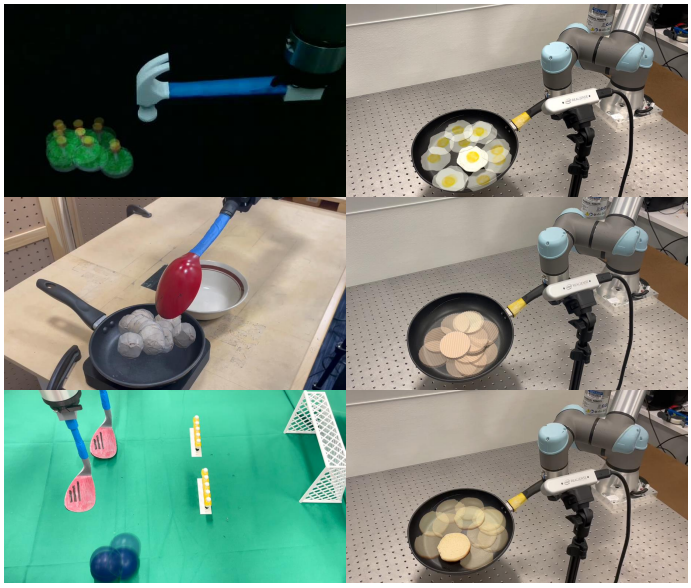


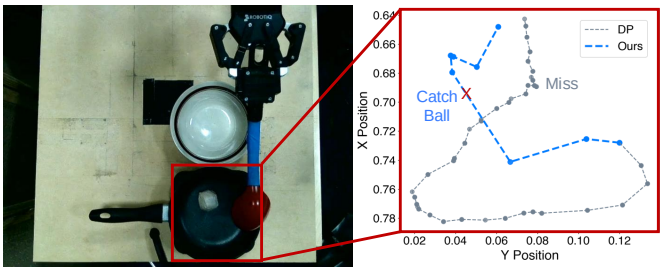
Fig. 7: **Initial States for All Evaluation Episodes.** All methods are evaluated using the same set of manually defined initial states, overlaid in the image. These states ensure diverse variations to test the policy's spatial generalization capabilities.

2) Implementation Details: **Hardware Design** We designed two fast tool changers compatible with robots using the ISO 9409-1-50-4-M6 flange, as shown in Figure 6. The left design utilizes a screw mechanism to accommodate general tools, while the right design employs clips for tools with specific mounting shapes.

Tool Pose Estimation We use Polycam to scan the tool and obtain its mesh. The mesh is later feed into FoundationPose [26] for 6D pose estimation.

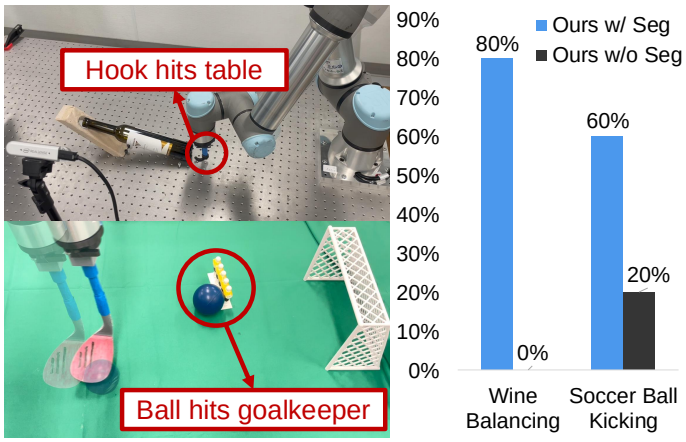
C. Additional Experimental Results

Policy Execution Trajectory Comparison: Our framework produces faster, smoother, and more natural trajectories compared to traditional approaches, as shown by the end-effector



(a) Initial Setup (b) EEF Trajectory Comparison

Fig. 8: **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.



(a) Typical Failures Without Embodiment Segmentation (b) Quantitative Results

Fig. 9: **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.

(EEF) XY trajectory for the meatball scooping task in Figure 8. Figure 8(a) shows the task setup, and Figure 8(b) compares our policy rollout with a baseline trained on robot-collected data. Our trajectory is significantly smoother, with 10× fewer waypoints, resulting in more fluid execution, reduced cumulative errors, and improved sample efficiency, thereby mitigating the distribution shifts commonly observed in behavior cloning. In contrast, the baseline exhibits excessive waypoints and discontinuous motions that hinder precise task execution.

Effects of Embodiment Segmentation: Embodiment Segmentation masks the agent's embodiments during data col-

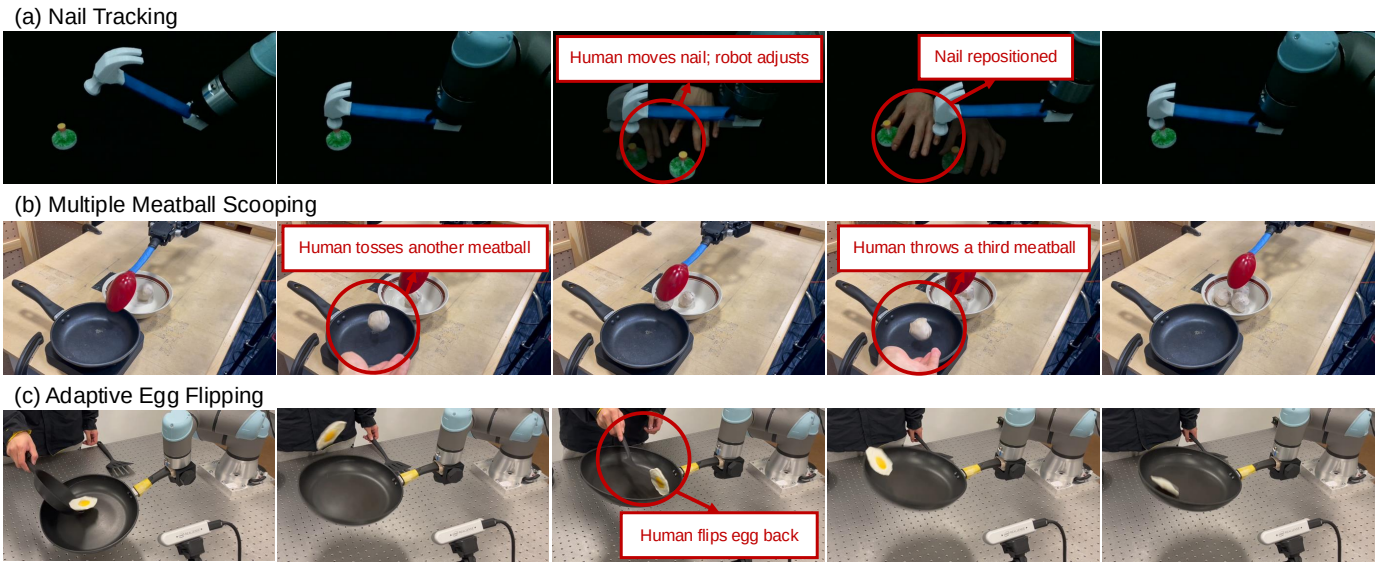
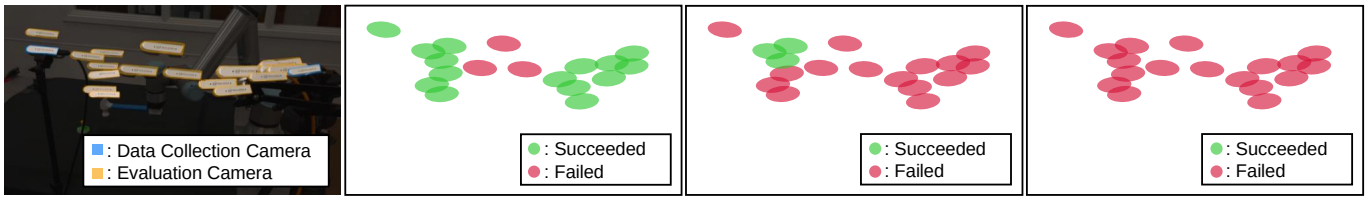


Fig. 10: **Human Perturbation Robustness.** The robot handles human-induced perturbations across three tasks: (1) In nail hammering, it tracked a manually moved nail; (2) In meatball scooping, it located and scooped new meatballs thrown mid-task; and (3) In egg flipping, it recovered the egg after human repositioning.



(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. 11: **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).

lection and policy deployment, ensuring visually consistent scenes and reducing the training-deployment visual gap. Embodiment Segmentation significantly improves policy performance, as shown in Figure 9. Figure 9(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 9(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.

Effects of Random Cropping and View Augmentation: Our experiments show that random cropping (RC) and view augmentation (VA) together enhance policy robustness to camera pose variations. RC improves resilience to minor perturbations such as small movements or shaking, while VA exposes the model to a broader distribution of viewpoints during training. We evaluated these techniques on the nail hammering task (Figure 11), comparing three models: one trained with both RC and VA, one with RC only, and one without either. The combined use of RC and VA significantly expands the range of camera configurations under which the policy can successfully operate.

Benefits of Tool-Based Action Representation in Task Space: Using the tool pose in the camera frame works with a static camera but fails under camera movement due to unreliable real-time tracking and incorrect end-effector positioning in the base frame. Similarly, representing actions in the base frame fails under base movement due to the assumption of a fixed base-to-workspace transform. In contrast, representing actions in task space is invariant to both camera and base movement, enabling robust execution even under large viewpoint shifts and base movements.

D. Detailed Analysis on 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).

1) Data Collection Efficiency: Our framework achieves significantly higher data collection throughput than traditional methods, enabling more demonstrations within the same time-frame. The improvement is driven by the natural and intuitive

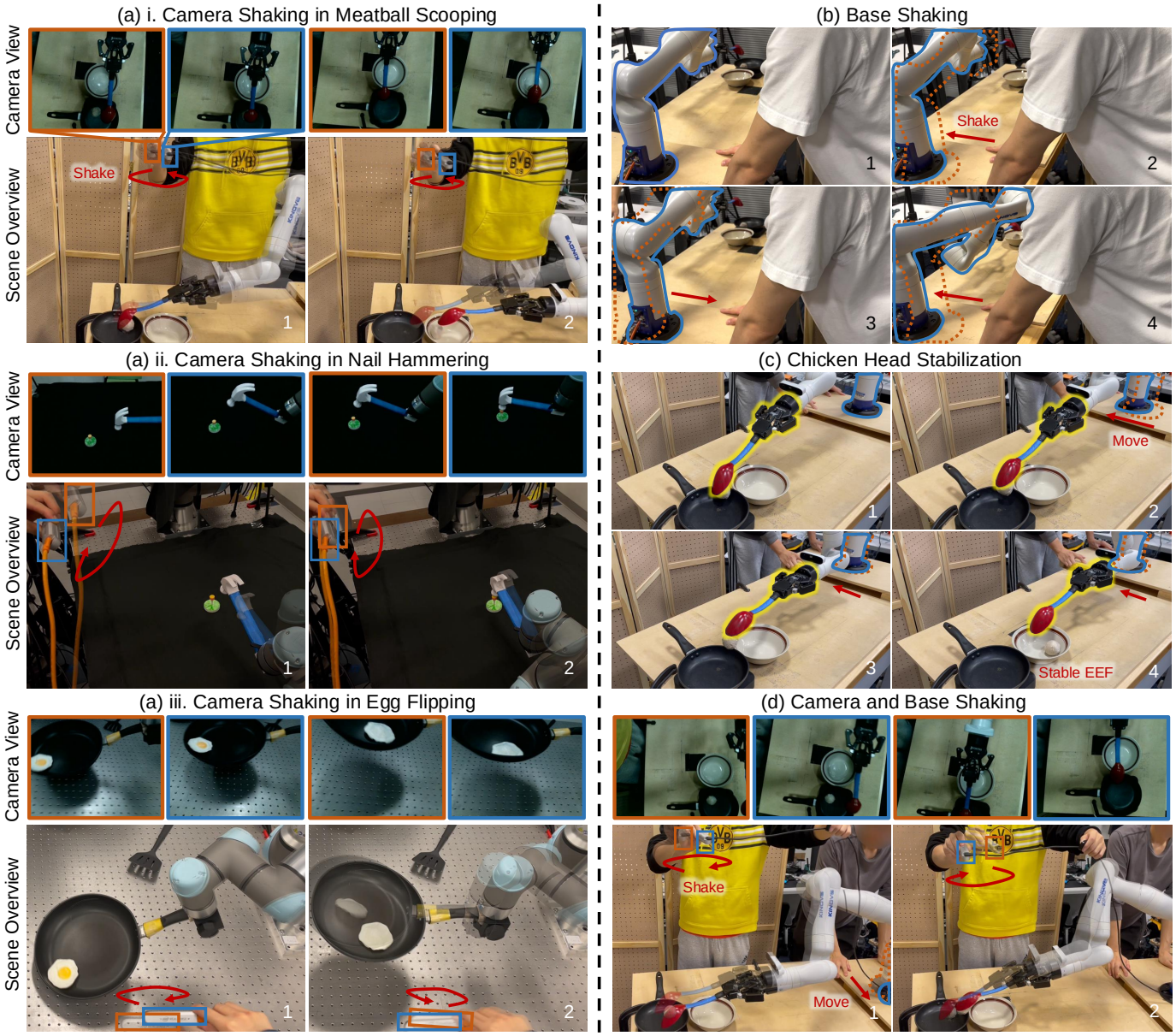


Fig. 12: **Robustness to Camera and Base Movement.** (a) Camera Pose Robustness: The policy demonstrated the ability to handle camera shaking across three tasks—meatball scooping, nail hammering, and pan flipping. The first row shows the camera view, while the second row provides a scene overview with 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.

efficiency of human manipulation, which ensures faster and more reliable task execution. Figure 13(a) highlights the superior manipulation capabilities of human hands, while Figure 14 quantifies the substantial 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 [8] 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 manipulation as a scalable solution for high-quality robot learning datasets.

2) Reliability: Figure 13(b) and Figure 13(c) illustrates typical failure cases with Gello, Spacemouse, and UMI [8], 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-

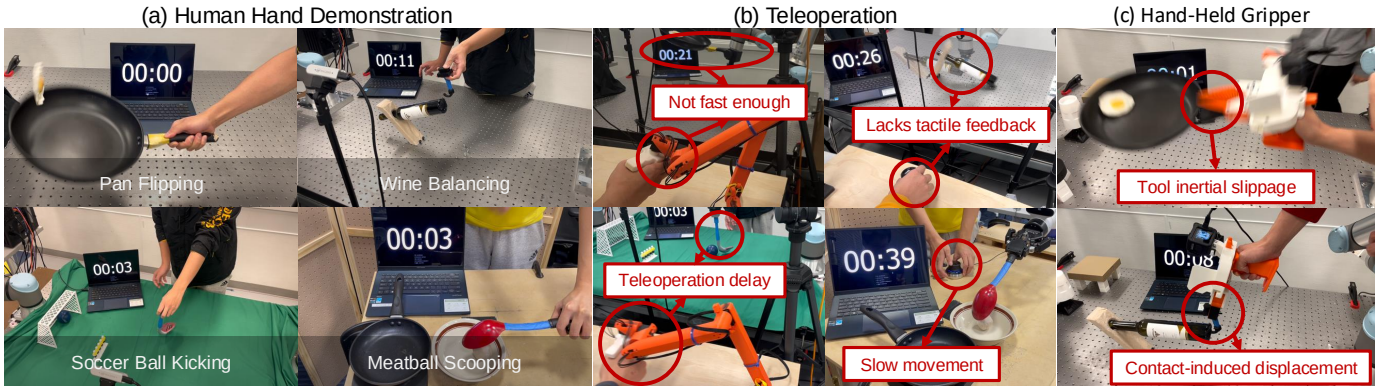


Fig. 13: **Data Collection Efficiency and Reliability.** (a) Human hands excel in manipulation tasks, leveraging natural and intuitive efficiency. (b) Failure cases for Gello and Spacemouse include insufficient speed, lack of tactile feedback during data collection, safety stops, collisions, teleoperation delays, and difficulty handling high-speed or complex tasks. (c) Failure cases for handheld grippers such as UMI [8], where issues arise from tool slippage due to inertia or displacement caused by contact forces.

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 [29]	$340	No	Yes	No	Moderate
UMI [8]	$371	No	Yes	**Yes**	Moderate
LEGATO [22]	$1060	No	Yes	**Yes**	Moderate
Spacemouse [9]	$169	**Yes**	Yes	**Yes**	Low
VR (Meta Quest 2 [17])	$300	**Yes**	Yes	No	Moderate
AR (Apple Vision Pro [1])	$3499	**Yes**	Yes	**Yes**	High
Gello [27]	$272	No	Yes	No	Moderate
Ours	**$0**	**Yes**	**No**	**Yes**	**Minimal**

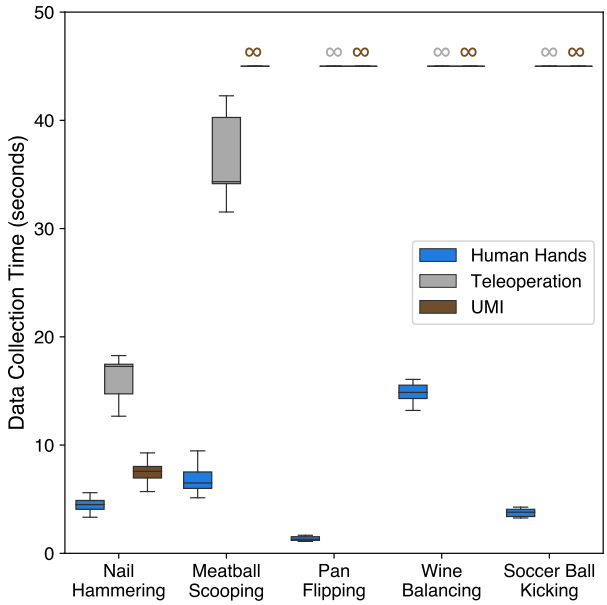


Fig. 14: **Quantitative Comparison of Data Collection Methods.** Human hands reduce data collection time by 73% for nail hammering and 81% for meatball scooping, while maintaining low variation. Teleoperation fails in dynamic and high-precision tasks. In nail hammering, human hands are 41% faster than UMI [8], which also struggles with dynamic and low-texture environments.

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 manipulation, our framework not only addresses these limitations but also provides a reliable and scalable solution for dynamic and precision-demanding tasks.

3) 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 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.