Tool-as-Interface: Learning Robot Tool Use from Human Play 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

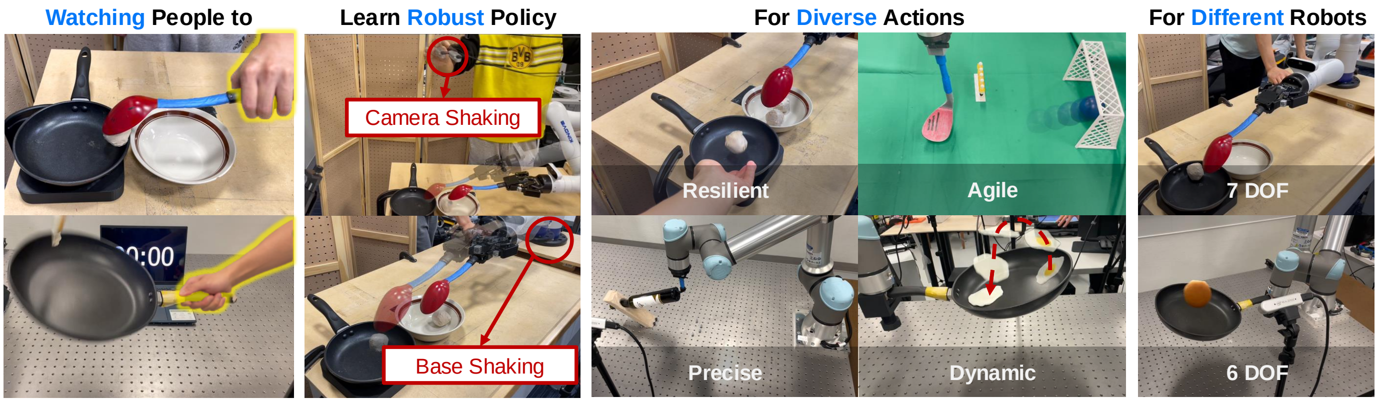


Fig. 1: **Tool-as-Interface.** We propose a scalable data collection and policy learning framework designed to transfer diverse, intuitive, and natural human play into effective visuomotor policies. The framework enables robots to learn robust policies that can operate effectively under challenging conditions,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

***Abstract*—Tool use is 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 play—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 play 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 [11], 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 [46, 28]. 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 [4, 32, 8, 24, 23]. 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 [15, 17, 18, 19]. 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 [21, 39, 6, 14, 19, 36, 52, 47, 26, 45, 31, 7] and hand-held grippers [44, 13, 37, 33, 30] 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 [53, 16, 49, 42, 34]. However, their reliance on direct access to robot hardware limits both practicality and scalability. Hand-held grippers [40, 11] 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 play — a natural, intuitive method through which humans interact with their environment during everyday activities without relying on external devices or specialized setups. Human play 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 play involves the spontaneous use of tools to interact with the environment. Human play 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 play. Key challenges include the embodiment gap and the reliance on single-view data, which limits the insights that can be drawn from human play data [41, 2, 25, 3, 43, 29, 38].

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 play 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 play 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 play 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 across a range of challenging real-world tool-use tasks, including nail hammering, meatball scooping, pan flipping, wine bottle balancing, and soccer ball kicking. Our method achieves a 71% improvement in success rate and reduces data collection time by 77% compared to diffusion policies trained on SpaceMouse [12] or Gello [49] data, with some tasks solvable only by our method. Our method also outperforms handheld grippers like UMI [11], 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. Our Robot Policy Learning Framework

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.

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 [35] 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 [20]. 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 [22], 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 to capture demonstration data. Then MASt3R processes the images to reconstruct a 3D point cloud without requiring camera extrinsics or intrinsics, and globally align 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 play 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 [9, 10]. Random cropping further improves the method's robustness, enabling the policy to generalize better to variations in visual inputs.

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 repre-

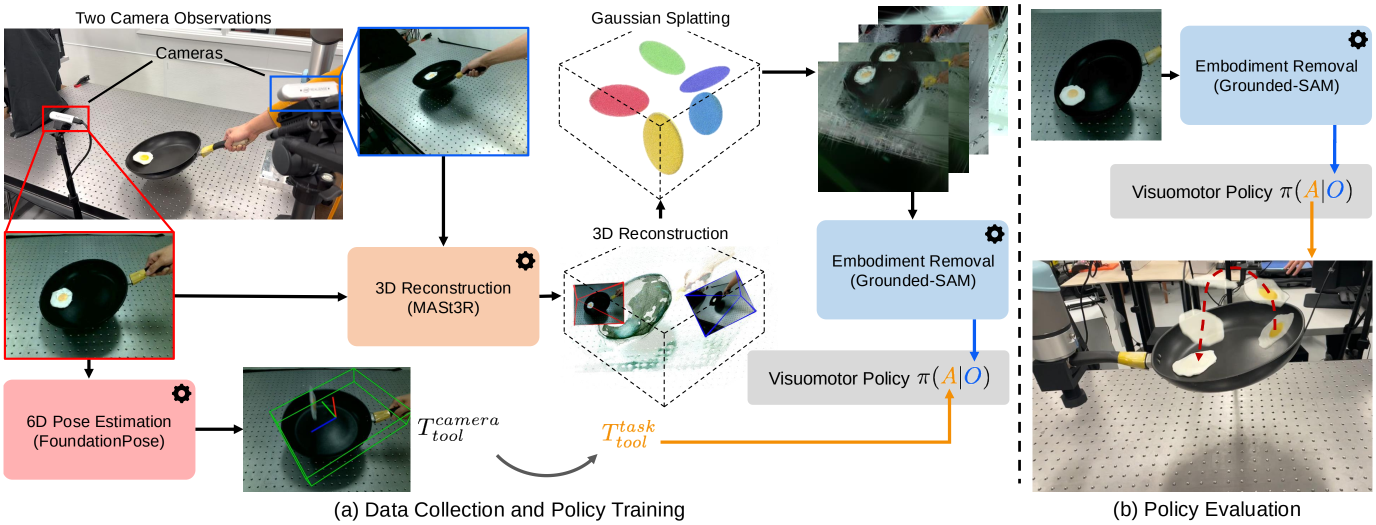


Fig. 2: **Policy Design.** Human play data was collected using two RGB cameras and processed through the foundation model MASt3R [22] to generate 3D reconstructions. Using 3D Gaussian splatting, we sampled novel views to augument the dataset. The human hand (embodiment) was segmented out from the images to create embodiment-agnostic observations, which serve as inputs to the policy. To label actions for policy training, a pose estimation model (FoundationPose [48] 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.

sentation focuses on the tool being manipulated, independent of human or robot morphology or camera pose.

Using a 6D pose estimation model (e.g., FoundationPose [48]), 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 [9] as our policy representation to predict $T_{\text{tool}}^{\text{task}}$, trained using the ACCESS system [5]. 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.

III. 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 [11], 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.

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.

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.

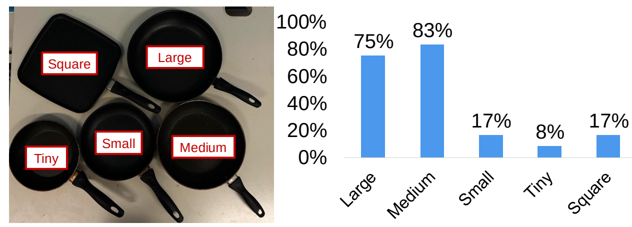
Benchmark	High-Precision	Extreme Dynamics	Using Extrinsic Dexterity	Contact-Rich	Overcoming Gravity
Task 1: Nail Hammering	✓	–	–	–	–
Task 2: Meatball Scooping	✓	–	✓	✓	–
Task 3: Pan Flipping (Egg, Bun, Patty)	–	✓	✓	✓	✓
Task 4: Wine Balancing	✓	–	✓	✓	✓
Task 5: Soccer Ball Kicking	–	–	–	✓	–

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 [11]	∼180 seconds (25 demos)	0/13
UMI	∼720 seconds (100 demos)	**13/13**
Ours	∼180 seconds (40 demos)	**13/13**



(a) Different Pans Used (b) Success Rate

Fig. 3: **Tool Generalization** (a) The tested pans. (b) Success rate across 12 testing trials.

C. Generalization

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

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. 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 3D Connexion. Spacemouse, 2023. URL https://3dconnexion.com/us/spacemouse/.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] P. Jonsson. Ericsson mobility report november 2024. In

Ericsson Mobility Report. Ericsson, 2024.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

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

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

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

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.