TeleMoMa: A Modular and Versatile Teleoperation System for Mobile Manipulation

Shivin Dass[1], Wensi Ai[2], Yuqian Jiang[1], Samik Singh[1], Jiaheng Hu[1],
Ruohan Zhang[2], Peter Stone[1,3], Ben Abbatematteo[1], Roberto Martín-Martín[1]

***Abstract*— A critical bottleneck limiting imitation learning in robotics is the lack of data. This problem is more severe in mobile manipulation, where collecting demonstrations is harder than in stationary manipulation due to the lack of available and easy-to-use teleoperation interfaces. In this work, we demonstrate TeleMoMa, a general and modular interface for whole-body teleoperation of mobile manipulators. TeleMoMa unifies multiple human interfaces including RGB and depth cameras, virtual reality controllers, keyboard, joysticks, etc., and any combination thereof. We demonstrate the versatility of TeleMoMa by teleoperating several existing mobile manipulators — PAL Tiago++, Toyota HSR, and Fetch — in simulation and the real world. We demonstrate the quality of the demonstrations collected with TeleMoMa by training imitation learning policies for mobile manipulation tasks involving synchronized whole-body motion. With a user study we demonstrate the importance of TeleMoMa's modularity. For more information and video results, `robin-lab.cs.utexas.edu/telemoma-web/`.**

I. INTRODUCTION

A core goal of robotics is to build generalist robots capable of operating alongside humans in their environment. To this end, learning from human-collected robot demonstrations has shown promise in endowing robots with the capabilities to solve complex tasks [1], [2], [3], boosted recently by the advent of foundation models capable of learning from large amounts of data [4], [5]. While these models demonstrate an impressive understanding of the tasks [6], [7], [8], [9], [10], these successes have been largely limited to stationary manipulation. However, a large fraction of the tasks that we would like generalist robots to perform require a combination of manipulation and mobility: e.g., sweeping the floor requires moving the broom with both hands and walking around to reach the dirty spots.

One of the reasons why stationary manipulation has enjoyed the benefits of large models, while mobile manipulation has not, is due to the availability of large datasets of human-collected demonstrations [8], [11]. They were obtained due to the multiple existing and easy-to-use teleoperation frameworks for stationary manipulators [12], [13], [14], [15], [16]. For mobile manipulation, however, the existing stationary manipulation teleoperation systems are not sufficient, due to the additional degrees of freedom that the user has to control including mobility and possibly multiple arms.

Several teleoperation frameworks for mobile manipulation have been proposed in the past, with different capabilities and limitations. They either enable accurate control with specific (and often expensive) hardware like motion capture systems [17], [18], [19] or puppeteering interfaces [20], [21], or achieve scalability by overloading simple and available devices that work for stationary manipulators such as gamepads [16], virtual reality controllers [22], or mobile phones [23], [24], limiting the expressiveness of the demonstrations. Teleoperation based solely on vision [25], [15], [26] promises an available and accessible interface at the cost of accuracy and dexterity. Each device alone presents a tradeoff between accuracy and availability, versatility and expressiveness, and as a result, no single device enables scalable, expressive teleoperation for all mobile manipulators.

Inspired by the complementary capabilities of several of the human interfaces for teleoperation, we introduce TeleMoMa (**Tele**operation for **Mo**bile **Ma**nipulation). TeleMoMa enables users to teleoperate different mobile manipulators with a variety of human interfaces or combinations thereof, in simulation or the real world, providing users the means to select the combination that best fits their teleoperation needs. We evaluate the benefits of modularity in a user study and find that a hybrid vision-VR interface is an efficient and natural mode of teleoperation. We also successfully trained several imitation learning policies on the data collected using TeleMoMa, indicating that TeleMoMa can collect high quality demonstrations. Further, in Appendix C, we perform extensive evaluations on remote teleoperation under network delays and comparison of teleoperation across different embodiments and, sim and real.

II. RELATED WORK

Successes in learning from large collections of human demonstrations has been limited to stationary manipulators [8], [6], [7] or simple mobile manipulation tasks like pick and place that do not require coordination between base and arm motion [27], [10]. This is in part due to the lack of accessible and intuitive ways to collect demonstrations for mobile robots. Recently, some methods have tried to address this using specialized hardware, such as motion capture systems [17], [18], [28], [19], exoskeletons [20], [29], [30], [31] and more sophisticated human-computer interfaces [32], [33]. On the other hand, several works borrow from successful teleoperation interfaces in stationary manipulation, using interfaces such as VR [34], [22], [35], [36], [37], [38], [39], kinesthetic teaching [29], visual motion tracking [25], keyboard and mouse [40] and mobile phones [24], by modifying them to enable the control of mobile manipulators. Although these interfaces are accessible, they lack the granularity

[1]The University of Texas at Austin [2]Stanford University [3]Sony AI

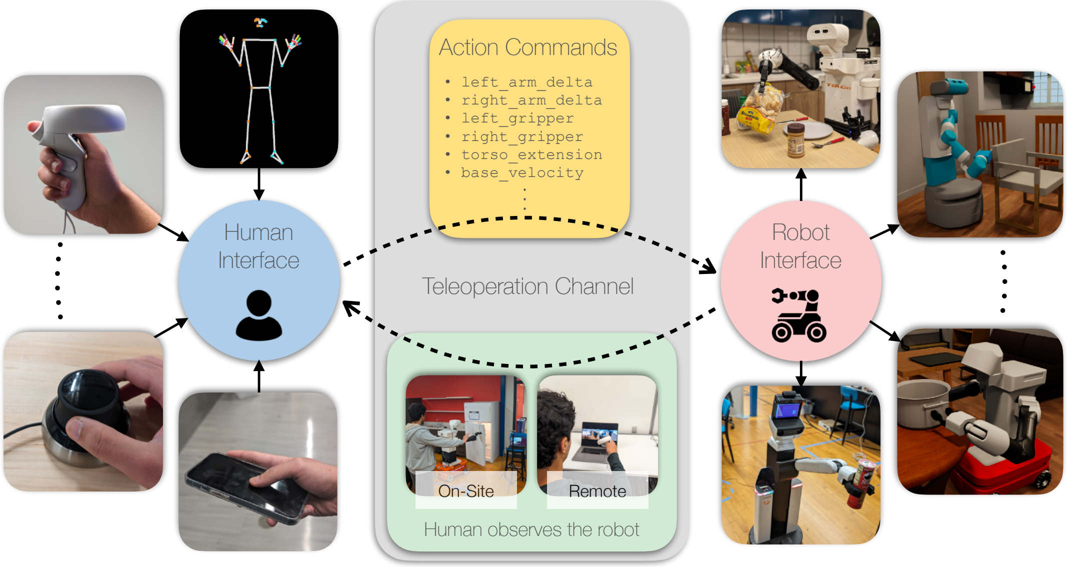


Fig. 1: **TeleMoMa System.** TeleMoMa consists of three components: the *Human Interface* acquires commands from the human using different input devices; the *Teleoperation Channel* defines the action command structure between the human and the robot interfaces, and, possibly, closes the loop with observations from the robot; and the *Robot Interface* implements a robot-specific mapping of actions to low-level robot commands. This architecture enables modularity and versatility – combining multiple devices to achieve intuitive whole-body teleoperation for multiple tasks and robots.

necessary to coordinate all degrees of freedom of a mobile robot for a true mobile manipulation task. For a more detailed overview of related works, see Appendix A.

III. TELEMOMA SYSTEM

TeleMoMa is a teleoperation system for mobile manipulators. It is generally composed of a *Teleoperation Channel* that defines the communication between a *Human Interface* and a *Robot Interface* (Fig. 1). The *Human Interface* acquires human inputs across different teleoperation modes such as vision, VR, spacemouse, keyboard, and mobile phones, or their combinations, and maps them to a general mobile manipulation action command structure provided by the *Teleoperation Channel* that includes fields such as base, arm, gripper, and torso motion. Multiple input devices can be combined through our *Human Interface* to acquire the action commands in the best suited manner for a task. The *Teleoperation Channel* hands over the action commands to the *Robot Interface*, a robot-specific module that maps the actions to robot motor commands. In the following, we provide additional information about the three components of TeleMoMa.

A. Human Interface

The *Human Interface* is responsible for processing the captured data from various teleoperation input devices and mapping them to a common action command structure. For each input device, the data is processed independently by a device-specific parser that maps the signals from the input modality (keyboard strokes, motion of a VR controller, location of human skeleton keypoints on an image, . . .) into elements of the teleoperation channel's action command. TeleMoMa supports input modalities such as vision, keyboard, spacemouse, VR (Oculus Quest and HTC Vive) and mobile phones. We explain the implementation specific details of each of these interfaces in Appendix B.

B. Teleoperation Channel

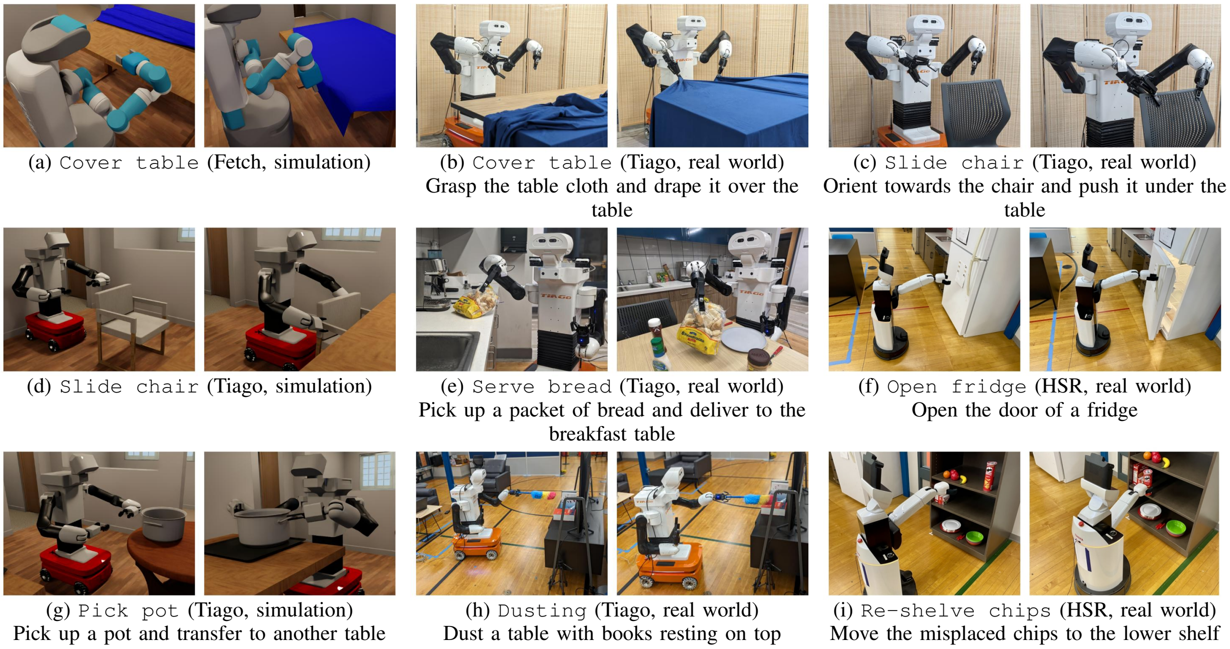
The *Teleoperation Channel* defines how the *Human Interface* communicates with the *Robot Interface*, and is the key to TeleMoMa's generality and modularity. Specifically, the *Teleoperation Channel* defines an action command structure that serves as a bridge between the human and the robot and the way the active human interfaces populate the entries of this structure.

During deployment, users can specify what input modality they want to use to control each part of the robot's embodiment including left and right arms and hands, torso, and base. The *Teleoperation Channel* automatically manages the action assignment based on the user specification, and consolidates the possible missing elements of the action commands due to differences in hardware frequency or network delays.

Finally, the *Teleoperation Channel* also defines the mechanism by which humans *close the loop* with the robot and observe the execution of the action commands. We consider two methods of observation: on-site and remote. When on-site, the human directly observes the robot executing the action commands. When remote, the *Teleoperation Channel* communicates the images from the onboard sensors of the robot to the human interface to be displayed for the human, enabling teleoperation from a different location.

C. Robot Interface

The *Robot Interface* is a robot-specific module that maps the commands obtained from the *Human Interface* to the motor commands on the robot. The specific controllers used to compute the torques are not part of the TeleMoMa system but they are necessary to map the action commands obtained



(a) `Cover table` (Fetch, simulation)

(b) `Cover table` (Tiago, real world)
Grasp the table cloth and drape it over the table

(c) `Slide chair` (Tiago, real world)
Orient towards the chair and push it under the table

(d) `Slide chair` (Tiago, simulation)

(e) `Serve bread` (Tiago, real world)
Pick up a packet of bread and deliver to the breakfast table

(f) `Open fridge` (HSR, real world)
Open the door of a fridge

(g) `Pick pot` (Tiago, simulation)
Pick up a pot and transfer to another table

(h) `Dusting` (Tiago, real world)
Dust a table with books resting on top

(i) `Re-shelve chips` (HSR, real world)
Move the misplaced chips to the lower shelf

Fig. 2: **Tasks in our evaluation of TeleMoMa**. Shown above is the initial and goal state of each task.

from the *Teleoperation Channel* into low-level commands. We do not deem our requirements for the robot platforms too high: the robot should provide some controllers to move either the end-effector(s) and the base in Cartesian space, the joints, or combinations of both.

The action command structure in TeleMoMa relayed to the *Robot Interface* can either contain values in task-space (end-effector Cartesian relative motion), joint space (e.g., torso commands or motion to other joints) and/or velocities (e.g., base commands), or different combinations of those, as specified by the user during deployment. The *Robot Interface* processes these commands based on the particular robot embodiment, filters out the unusable action components (such as left hand commands for a single-armed robot like Fetch), and maps the rest to the robot using the preferred choices of controllers such as operational space control [41] to control one task frame, or whole-body control [42], [43] to command the entire robot jointly.

IV. EXPERIMENTS

In our experiments we seek to answer the following questions: (1) What are the benefits of TeleMoMa's modularity? (Sec. IV-A) (2) Can TeleMoMa collect high-quality data for imitation learning? (Sec. IV-B)

Additional experiments on remote teleoperation under network delays and comparison of teleoperation across different embodiments and, sim and real are explored in Appendix C.

A. User Study

To assess the performance of different teleoperation modalities in the TeleMoMa framework, we performed a user study with the PAL Tiago++ robot. We compared three teleoperation modalities: *VR*, in which the user controls the robot's arms with the Oculus controllers and the base and torso with the controller joysticks; *Vision*, in which the user's pose is tracked with an RGB-D camera to control the arms, torso and base motion; and *VR+Vision* combining both modalities, in which the robot's arms are controlled using the Oculus controllers and the base and torso motion is controlled via human pose tracking from RGB-D data.

We compared the three modalities (*VR*, *Vision*, *VR* + *Vision*) to assess the completion time in two tasks: `cover table` (Fig. 2(b)), in which the robot must grasp a tablecloth with both hands and drape it over a table, and `dusting` (Fig. 2(h)), in which the robot must dust a table with books resting on top. Both tasks, but especially the `dusting` task, benefit from the simultaneous motion of base and arm(s), i.e., whole-body motion, as enabled by TeleMoMa since the robot is required to navigate around the desk while periodically moving the hands to clear out any dust.

We recruited 12 participants with varying levels of teleoperation experience. Each user was given the same instructions and a brief practice period with each modality. The order in which users received the devices was randomized. The completion times for successful trials are provided in Fig. 3. The only failures observed occurred with the *Vision* modality (3 fails out of 12 `dusting` trials) due to noise and inaccuracies in the pose tracking. We observe that in the `cover table` task, performance is comparable across teleoperation modalities. However, in the `dusting` task, pure *VR* is generally slower than *VR* + *Vision* or *Vision* alone due to the lack of intuitive whole-body teleoperation: because moving the base requires using the joysticks on the controllers, users tended to only move the arm or the base one at a given time. The results indicate that on their own, both *VR* and *Vision* present drawbacks pertaining to their individual modalities, but when combined in the form

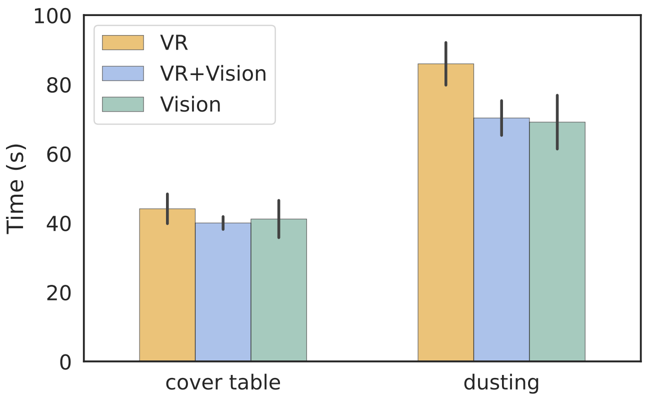


Fig. 3: **User Study: Completion Time.** Vision modalities outperform only-VR for the more challenging `dusting` task. Error bars denote the standard error of the mean.

TABLE I: Performance between IL policies trained with RGB vs. RGBD images as inputs. Successes measured over 10 rollouts.

	Cover Table		**Slide Chair**		**Serve Bread**	
Modality	RGB	RGB-D	RGB	RGB-D	RGB	RGB-D
BC	60	60	40	60	20	40
BC-RNN	70	**90**	50	**80**	30	**70**

TABLE II: IL Policy performance scale with data. Successes measured over 10 rollouts.

	Cover Table		**Slide Chair**		**Serve Bread**	
Fraction of data	50%	100%	50%	100%	50%	100%
BC	60	60	40	60	30	40
BC-RNN	60	**90**	70	**80**	40	**70**

of *VR + Vision*, TeleMoMa can overcome their individual drawbacks to enable an improved teleoperation experience. These results support empirically the importance of enabling multiple input modalities and their combination for teleoperation of mobile manipulators.

B. Imitation Learning with TeleMoMa's Data

To empirically evaluate the quality of data collected with TeleMoMa, we train several visuomotor policies with behavioral cloning [44] on a Tiago++ robot (*real*). We consider three diverse mobile manipulation tasks:

- `cover table`: Similar to the one described in Sec. IV-A, the tasks involves bimanual grasping of a tablecloth and draping it over a table (Fig. 2(b)).
- `slide chair`: A bimanual task, that requires the robot to navigate and align itself behind a chair, grasp it, and push the chair under a table (Fig. 2(c)).
- `serve bread`: In a real kitchen setting, the robot is required to navigate to the kitchen counter, pick a bag of bread, and deliver it to the breakfast table (Fig. 2(e)).

We collected 50 demonstrations each for `slide chair` and `serve bread` tasks and 100 demonstrations for `cover table` task using the combined *VR + Vision* interface of TeleMoMa. Additional demonstrations in the `cover table` were necessary to allow the policies to learn the necessary accurate grasps on the cloth.

Policy Architecture, Observations and Actions. We used a feed-forward MLP (BC) and a recurrent LSTM based network (BC-RNN) [3] with a sequence length of 10. The inputs to all policies included RGB-D images obtained from two realsense cameras attached on each shoulder of the robot, end-effector poses of the hands, gripper state, and the change in the mobile base pose obtained from the odometry of the robot. The policies output a 17-dimensional action space: 6D Cartesian deltas and a gripper command for each of the hands, and linear and angular velocities for the base.

Comparing Input Modalities. To analyze the importance of depth sensing in learning mobile manipulation tasks, we train two sets of policies: the first set was trained exclusively on RGB observations, while the second combined RGB and Depth. The performance of the two sets of policies for each of the tasks is summarized in Table I. Our analysis reveals a consistent trend: irrespective of the policy architecture, the inclusion of depth information markedly enhances performance across all tasks. Qualitatively, we observe that policies trained using depth can position better, significantly improving the efficacy of subsequent arm actions. These findings suggest that depth information is a crucial component for the development of effective mobile manipulation policies.

Performance with Different Amounts of Data. To investigate how data volume influences policy performance, we experimented with two distinct policy groups: the first group was trained using the complete dataset we gathered for each task, while the second group utilized only 50% of these collected demonstrations. The results are summarized in Table II; we observe that policies trained with the full dataset consistently outperform those trained with half the data, demonstrating the importance of dataset size in imitation learning, especially in this low-data regime. We additionally notice that BC-RNN strictly outperforms regular BC in all tasks, demonstrating the significance of temporal dependencies for learning mobile manipulation tasks.

In general, the above experiments provide compelling evidence that IL policies trained with data collected using TeleMoMa can reliably perform complex mobile manipulation tasks, thus indicating that TeleMoMa can facilitate high-quality data collection for imitation learning. We demonstrate more imitation results in the sim environment in Appendix C.

V. CONCLUSIONS

In closing, we have demonstrated TeleMoMa, a general, modular, accessible teleoperation system that enables collection of high-quality expert demonstration data for a variety of complex and novel mobile manipulation tasks. We showed TeleMoMa's generality by teleoperating multiple different robots in simulation and reality, and conducted user studies to verify the usability of the system's various modalities. We hope that our system lowers the barrier of entry for researchers to collect high-quality demonstrations for mobile manipulation, and helps unlock new mobile manipulation capabilities.

References

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[2] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[3] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *arXiv preprint arXiv:2108.03298*, 2021.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[5] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, *et al.*, "Foundation models in robotics: Applications, challenges, and the future," *arXiv preprint arXiv:2312.07843*, 2023.

[6] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, *et al.*, "Octo: An open-source generalist robot policy," 2023.

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

[8] O. X.-E. Collaboration, A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, A. Raffin, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Ichter, C. Lu, C. Xu, C. Finn, C. Xu, C. Chi, C. Huang, C. Chan, C. Pan, C. Fu, C. Devin, D. Driess, D. Pathak, D. Shah, D. Büchler, D. Kalashnikov, D. Sadigh, E. Johns, F. Ceola, F. Xia, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Schiavi, H. Su, H.-S. Fang, H. Shi, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Kim, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Wu, J. Luo, J. Gu, J. Tan, J. Oh, J. Malik, J. Tompson, J. Yang, J. J. Lim, J. Silvério, J. Han, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Zhang, K. Majd, K. Rana, K. Srinivasan, L. Y. Chen, L. Pinto, L. Tan, L. Ott, L. Lee, M. Tomizuka, M. Du, M. Ahn, M. Zhang, M. Ding, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, P. R. Sanketi, P. Wohlhart, P. Xu, P. Sermanet, P. Sundaresan, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Moore, S. Bahl, S. Dass, S. Song, S. Xu, S. Haldar, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Dasari, S. Belkhale, T. Osa, T. Harada, T. Matsushima, T. Xiao, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Li, Y. Lu, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. hua Wu, Y. Tang, Y. Zhu, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Xu, and Z. J. Cui, "Open X-Embodiment: Robotic learning datasets and RT-X models," https://arxiv.org/abs/2310.08864, 2023.

[9] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[10] Y. Chebotar, Q. Vuong, K. Hausman, F. Xia, Y. Lu, A. Irpan, A. Kumar, T. Yu, A. Herzog, K. Pertsch, *et al.*, "Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions," in *Conference on Robot Learning*. PMLR, 2023, pp. 3909–3928.

[11] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[12] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Conference on Robot Learning*. PMLR, 2018, pp. 879–893.

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

[14] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," *arXiv preprint arXiv:2309.13037*, 2023.

[15] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *arXiv preprint arXiv:2307.04577*, 2023.

[16] S. Dass, K. Pertsch, H. Zhang, Y. Lee, J. J. Lim, and S. Nikolaidis, "Pato: Policy assisted teleoperation for scalable robot data collection," *arXiv preprint arXiv:2212.04708*, 2022.

[17] M. Arduengo, A. Arduengo, A. Colomé, J. Lobo-Prat, and C. Torras, "Human to robot whole-body motion transfer," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2021, pp. 299–305.

[18] C. Stanton, A. Bogdanovych, and E. Ratanasena, "Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning," in *Proc. Australasian Conference on Robotics and Automation*, vol. 8, 2012, p. 51.

[19] A. Setapen, M. Quinlan, and P. Stone, "Marionet: Motion acquisition for robots through iterative online evaluative training," in *Ninth International Conference on Autonomous Agents and Multiagent Systems - Agents Learning Interactively from Human Teachers Workshop (AAMAS - ALIHT)*, May 2010.

[20] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[21] A. Purushottam, C. Xu, Y. Jung, and J. Ramos, "Dynamic mobile manipulation via whole-body bilateral teleoperation of a wheeled humanoid," *IEEE Robotics and Automation Letters*, 2023.

[22] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu, "Deep imitation learning for humanoid loco-manipulation through human teleoperation," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2023, pp. 1–8.

[23] A. Tung, J. Wong, A. Mandlekar, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese, "Learning multi-arm manipulation through collaborative teleoperation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9212–9219.

[24] J. Wong, A. Tung, A. Kurenkov, A. Mandlekar, L. Fei-Fei, S. Savarese, and R. Martín-Martín, "Error-aware imitation learning from teleoperation data for mobile manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1367–1378.

[25] Z. Zhang, Y. Niu, Z. Yan, and S. Lin, "Real-time whole-body imitation by humanoid robots and task-oriented teleoperation using an analytical mapping method and quantitative evaluation," *Applied Sciences*, vol. 8, no. 10, p. 2005, 2018.

[26] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9164–9170.

[27] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

[28] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The kit bimanual manipulation dataset," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2021, pp. 499–506.

[29] T. Yang, Y. Jing, H. Wu, J. Xu, K. Sima, G. Chen, Q. Sima, and T. Kong, "Moma-force: Visual-force imitation for real-world mobile manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6847–6852.

[30] Y. Matsuura, K. Kawaharazuka, N. Hiraoka, K. Kojima, K. Okada, and M. Inaba, "Development of a whole-body work imitation learning system by a biped and bi-armed humanoid," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 10 374–10 381.

[31] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Low-cost exoskeletons for learning whole-arm manipulation in the wild," *arXiv preprint arXiv:2309.14975*, 2023.

[32] M. Schwarz, C. Lenz, R. Memmesheimer, B. Pätzold, A. Rochow, M. Schreiber, and S. Behnke, "Robust immersive telepresence and mobile telemanipulation: Nimbro wins ana avatar xprize finals," *arXiv preprint arXiv:2303.03297*, 2023.

[33] C. Lenz and S. Behnke, "Bimanual telemanipulation with force and haptic feedback through an anthropomorphic avatar system," *Robotics and Autonomous Systems*, vol. 161, p. 104338, 2023.

[34] B. R. Galarza, P. Ayala, S. Manzano, and M. V. Garcia, "Virtual reality teleoperation system for mobile robot manipulation," *Robotics*, vol. 12, no. 6, p. 163, 2023.

[35] L. Penco, K. Momose, S. McCrory, D. Anderson, N. Kitchel, D. Calvert, and R. J. Griffin, "Mixed reality teleoperation assistance for direct control of humanoids," *IEEE Robotics and Automation Letters*, 2024.

[36] A. Garcia-Garcia, P. Martinez-Gonzalez, S. Oprea, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Jover-Alvarez, "The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6790–6797.

[37] C. Gan, J. Schwartz, S. Alter, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, M. Sano, *et al.*, "Threedworld: A platform for interactive multi-modal physical simulation," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[38] P. Martinez-Gonzalez, S. Oprea, A. Garcia-Garcia, A. Jover-Alvarez, S. Orts-Escolano, and J. Garcia-Rodriguez, "Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation," *Virtual Reality*, vol. 24, pp. 271–288, 2020.

[39] G. Kazhoyan, A. Hawkin, S. Koralewski, A. Haidu, and M. Beetz, "Learning motion parameterizations of mobile pick and place actions from observing humans in virtual environments," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9736–9743.

[40] E. Ratner, B. Cohen, M. Phillips, and M. Likhachev, "A web-based infrastructure for recording user demonstrations of mobile manipulation tasks," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 5523–5530.

[41] "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 1, pp. 43–53, 1987.

[42] "Whole-body dynamic behavior and control of human-like robots," *International Journal of Humanoid Robotics*, vol. 1, no. 01, pp. 29–43, 2004.

[43] N. Mansard, O. Stasse, P. Evrard, and A. Kheddar, "A versatile generalized inverted kinematics implementation for collaborative working humanoid robots: The stack of tasks," in *2009 International conference on advanced robotics*. IEEE, 2009, pp. 1–6.

[44] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in cognitive sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[45] J. Vertut and P. Coiffet, *Teleoperations and robotics: evolution and development*. Prentice-Hall, Inc., 1986.

[46] B. Siciliano, O. Khatib, and T. Kröger, *Springer handbook of robotics*. Springer, 2008, vol. 200.

[47] A. Mandlekar, J. Booher, M. Spero, A. Tung, A. Gupta, Y. Zhu, A. Garg, S. Savarese, and L. Fei-Fei, "Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1048–1055.

[48] R. Hoque, L. Y. Chen, S. Sharma, K. Dharmarajan, B. Thananjeyan, P. Abbeel, and K. Goldberg, "Fleet-dagger: Interactive robot fleet learning with scalable human supervision," in *Conference on Robot Learning*. PMLR, 2023, pp. 368–380.

[49] D. Ryu, J.-B. Song, C. Cho, S. Kang, and M. Kim, "Development of a six dof haptic master for teleoperation of a mobile manipulator," *Mechatronics*, vol. 20, no. 2, pp. 181–191, 2010.

[50] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5628–5635.

[51] D. Whitney, E. Rosen, D. Ullman, E. Phillips, and S. Tellex, "Ros reality: A virtual reality framework using consumer-grade hardware for ros-enabled robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[52] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3758–3765.

[53] J. DelPreto, J. I. Lipton, L. Sanneman, A. J. Fay, C. Fourie, C. Choi, and D. Rus, "Helping robots learn: a human-robot master-apprentice model using demonstrations via virtual reality teleoperation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 226–10 233.

[54] X. Gao, R. Gong, T. Shu, X. Xie, S. Wang, and S.-C. Zhu, "Vrkitchen: an interactive 3d virtual environment for task-oriented learning," *arXiv preprint arXiv:1903.05757*, 2019.

[55] J. I. Lipton, A. J. Fay, and D. Rus, "Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 179–186, 2017.

[56] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube," *arXiv preprint arXiv:2202.10448*, 2022.

[57] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, Sept. 2013. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364913495721

[58] B. Wang, Z. Li, W. Ye, and Q. Xie, "Development of human-machine interface for teleoperation of a mobile manipulator," *International Journal of Control, Automation and Systems*, vol. 10, pp. 1225–1231, 2012.

[59] L. Fritsche, F. Unverzag, J. Peters, and R. Calandra, "First-person teleoperation of a humanoid robot," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 997–1002.

[60] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, vol. 2019, 2019.

APPENDIX

A. Extended Related Work

Teleoperation is almost as old as the field of robotics itself [45], with early manipulators being controlled in kinematically identical master-slave systems [46] similar to the very recent Mobile ALOHA [20]. More recently, teleoperation has emerged as a critical means of data collection for imitation learning methods [1], [2], as the ability to quickly collect large scale robotic data has become paramount for training large capacity behavior models [47], [48], [16]. Many teleoperation modalities have been proposed to address these challenges, including kinesthetic teaching, joysticks, virtual reality, mobile phones, RGB cameras, exoskeletons, and motion capture.

Each modality has its benefits and shortcomings. Joysticks (e.g. the SpaceMouse) offer intuitive control of a robot's end-effector(s), but fail to enable joint control or navigation [49]. Virtual reality enables users to perform tasks from the robot's perspective, but is limited by individual tolerance to motion sickness and does not naturally enable simultaneous locomotion and manipulation [50], [51], [52], [53], [54], [55]. Mobile phones offer scalable data collection, but provide a very limited interface, failing to naturally support joint control or base motion [12], [47]. RGB cameras have been explored as an accessible, scalable medium with limited mobility and range of motion [26], [15], [56]. Exoskeletons and master-slave devices enable dexterous control but are typically platform-specific and costly [31], [13], [20], [21], and do not naturally provide a way to coordinate base and arm motion. Motion capture similarly enables high-quality data collection, but is costly and difficult to scale [17], [19], [18]. Kinesthetic teaching was the predominant teleoperation paradigm for imitation learning for many years [1], [2], [57], but fails to enable more complicated bimanual or mobile manipulation tasks. Some works explore the combination of different modalities [58], [59] but fail to be sufficiently general and extensible. Thus, despite the plethora of available options, there remains a need for a teleoperation system capable of adapting to the needs of mobile manipulation in a scalable, accessible way.

We summarize the main features of TeleMoMa and contrast it with related systems in Table III. We compare across two primary dimensions: the teleoperation modalities provided, and the robot capabilities enabled. TeleMoMa is the only teleoperation system to provide modularity and enable the flexible combination of multiple input modalities.

B. Method Details

Following we describe how TeleMoMa facilitates the use of cameras (/vision), VR controllers, mobile phones, spacemouse, and keyboards as part of its *Human Interface* (Sec. III-A). We are also open-sourcing the code to the community to facilitate plug-and-play teleoperation for mobile manipulators to improve data collection efficiency.

1) Vision-Based Human Interface: TeleMoMa offers a unique vision-based pipeline for the whole-body teleoperation of a mobile manipulator using a single RGB-D camera. We use MediaPipe [60], a lightweight RGB-based model that executes in real-time for body pose and hand keypoint detection. Our proposed human interface uses the position and rotation of the hips to control the movement of the base of the mobile manipulator. Since the model only provides the relative depth of the keypoints to the center of the hip and not the absolute depth, we use the depth channel of an RGB-D camera to obtain the absolute values. The hand keypoints are mapped to the end-effector of the robot based on the position and orientation of the palm with respect to the hip. We compute the per-frame relative pose displacement in Cartesian space of the hands and send them in the teleoperation channel's action command as arm delta commands. Additionally, we use the distance between the center of the hips and ankles to command the robot height for robots with an actuated torso.

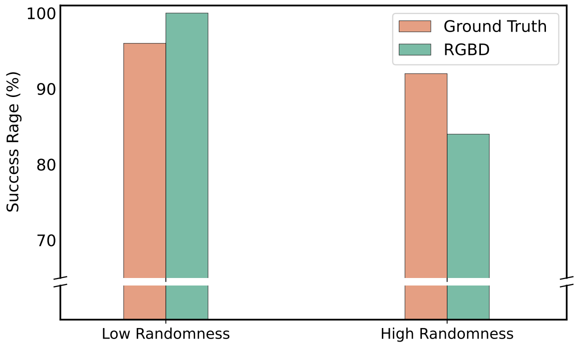


Fig. 4: **IL Results in Simulation.** Policy with RGBD input yields comparable performance to policy with ground truth chair positions as input.

2) Virtual Reality Controllers: TeleMoMa supports Oculus Quest and HTC Vive virtual reality hardware devices as inputs to the VR human interface. The controllers are tracked with respect to the headset for Oculus and with respect to the lighthouse for HTC Vive. Similar to [22], the tracked hand poses in Cartesian space are used to command the end-effector in the task-space. As in the vision-based interface, we compute the per-frame relative pose displacement of hands and use them in the teleoperation channel's action command. The joysticks integrated in the VR are used to command the velocities of the mobile base and also control the torso extension.

3) Mobile Phone: We use an app using the ARKit development kit to track the position and orientation of the mobile phone, which sends commands over the network. Similar to *Virtual Reality Controller*, the end effector is commanded in the task space and the relative pose displacement per frame of the mobile phone is calculated and mapped to the robot end effector. The gripper is controlled by dedicated buttons in the mobile app. Additionally, simultaneous control of left and right arms can be facilitated if two mobile phones are running the app, each phone controlling one of the arms. Mobile phones currently don't support navigation capabilities, but can be combined with other modalities such as the *Vision-*

TABLE III: Comparison of Existing Mobile Manipulation Teleoperation Systems

	Teleoperation Support			**Robot Support**					
	Cost / Accessibility	Modular	Modality	Bimanual	Height Control	Whole-Body Teleop	Robot Agnostic	Action Space	Domain
[17]	$$	×	Mocap	×	✓	✓	✓	EE Pose / Base Vel.	Real
MoMaRT [24]	$	×	Phone	×	×	×	✓	EE Pose / Base Vel.	Sim
MOMA-Force [29]	$$$	×	Kinesthetic	×	×	×	✓	EE Pose and Wrench	Real
SATYRR [21]	$$$	×	Puppeteer	✓	×	✓	×	Joint Pos. / Base Vel.	Real
TRILL [22]	$	×	VR	✓	×	✓	✓	EE Poses / Gait	Sim&Real
Mobile ALOHA [20]	$$$	×	Puppeteer	✓	×	✓	×	Joint Pos. / Base Vel.	Real
TeleMoMa	$	✓	*	✓	✓	✓	✓	EE Poses / Base Vel. / Joint Pos.	Sim&Real

based Human Interface to facilitate mobile base movements.

4) Spacemouse: Spacemouse has only 6-degrees of freedom, which is why we use mode switching, and control each part of the robot independently. The users can switch modes by pressing one of the side buttons of the spacemouse and switch between controlling left arm, right arm, base and torso. Two spacemouse' can also be used simultaneously for controlling each of the arms and minimizing the mode switching. The displacement of the spacemouse in each of the 6 degrees of freedom is tracked and sent as the delta commands to control the arms. For the base and torso, only the required displacements are used to send commands, while the remaining ones are discarded. The gripper can be toggled by pressing the remaining side button when the spacemouse mode is controlling the corresponding arm. Spacemouse gains significantly from modularity offered by TeleMoMa, by minimizing mode switching thus gaining more fluid control of the robot.

5) Keyboard: Keyboard presses are asynchronously read by the device listeners and each key is mapped to a single DoF of the mobile manipulator. Each key increases / decreases one of the DoFs in the Cartesian space by some preset amount. This results in a large number of keys that the teleoperator has to remember for controlling the robot. Instead, using a smaller set of keys for controlling for instance, just the base, while controlling arms with something more intuitive such as the spacemouse can drastically improve the teleoperation experience on both the interfaces, minimizing the mode switching in case of spacemouse, and reducing the number of keys to keep track of on the keyboard.

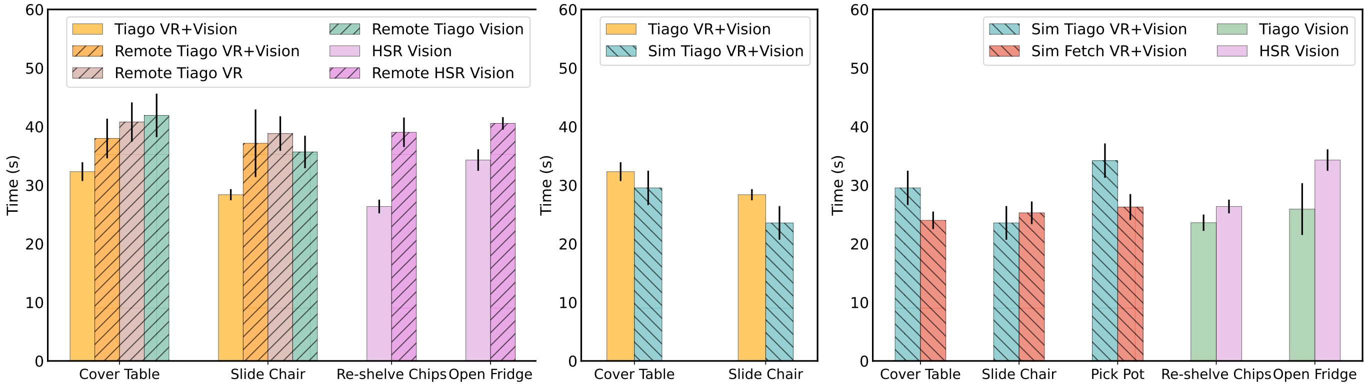
C. Additional Experiments

Imitation Results in Simulation. We show the imitation results of the `slide chair` (Fig. 2(d)) task in simulation here. We collected 100 demos in OmniGibson, and trained 2 policies using BC with different input observations: one with RGB-D image from the head camera, and the other with oracle chair positions in both world frame and robot base frame from the simulation environment. Robot proprioception, including end effector poses for two arms in base frame, and the base position and velocity in world frame, are also provided as observation input. We evaluated the policy on two task configurations: first with low randomness, where the chair position is uniformly sampled within 0.2 meters parallel to the robot, and second with high randomness, where the sampling interval is 1 meters. Each policy is evaluated with 25 rollouts under these conditions.

The results are shown in Fig. 4. We observed that, the performance of policies under high randomness is worse than under low randomness, which is expected because of the increased difficulty. We additionally observe that in both low and high randomness settings, policy trained with RGB-D input performs comparable to the one trained with ground truth chair positions, indicating that the policies are able to extract meaningful environment specific details from images and depth. Qualitatively, we observe that the causes of failure include misalignment between the robot and the chair, slippage of robot grippers, and knocking over the chair due to the application of excessive force.

Remote Teleoperation. TeleMoMa's architecture allows a remote demonstrator to control the robot from a client computer connected over the internet. Instead of watching the robot on-site, the demonstrator is provided with camera streams transmitted by the teleoperation channel from the robot's onboard sensors. To minimize communication delays, TeleMoMa 1) sends compressed sensor images from the robot and decompresses them on the client, and 2) in the case of a vision-based human interface, TeleMoMa processes the RGB-D images from the vision interface on the client side and only sends the action commands over the teleoperation channel. For other interfaces, the demonstrated action commands are directly sent to the TeleMoMa's robot interface.

We demonstrate the remote teleoperation capability of TeleMoMa on several combinations of robot hardware and user interfaces. To evaluate the effects of communication delays, we compare the task completion time between on-site and remote demonstrations using Tiago++ and Toyota HSR each on two different tasks. The `cover table` and the `slide chair` tasks are completed using Tiago++ with the on-site *VR* + *Vision* interface and three remote interfaces (*VR*, *Vision*, *VR* + *Vision*). The `re-shelve chips` task, in which the robot must move the misplaced chips to the lower shelf (Fig. 2(i)), and the `open fridge` task, in which the robot must open a fridge (Fig. 2(f)), are completed using HSR with the *Vision* interface. The demonstrations are provided by an expert user of each robot. The Wi-Fi



(a) On-Site vs. Remote Teleoperation (b) Real world vs. Simulation (c) Teleoperating different robot embodiments

Fig. 5: Completion times in different experiments with TeleMoMa. The bars indicate the mean and standard deviation of several trials (see text). *From left to right:* Comparing completion times for tasks performed on-site and remote, with HSR and Tiago; Completion times for real vs. simulated tasks with Tiago; Completion times for different robot embodiments on the same tasks in the real world and simulation. TeleMoMa allows for multiple tasks in simulation and the real world, with several embodiments

speed is about 100 Mbps as measured on the HSR. Fig. 5(a) shows the completion time in each modality averaged over 3 runs. We observe that remote human demonstrators have slower reaction times due to delays and limited resolutions of the camera streams, but TeleMoMa provides the capability to successfully complete the tasks under regular network conditions.

Comparing Different Embodiments and Sim vs. Real. In the final set of experiments, we seek to study how the domain (sim vs. real) and the type of robot (Tiago vs. HSR and Tiago-sim vs. Fetch-sim) influence the teleoperation behavior for the same tasks.

1) Sim vs. Real: Fig. 5(b) depicts the results of comparing completion time for `cover table` and `slide chair` tasks in simulation and real environment using a Tiago robot. We use sim time for simulation evaluation because of OmniGibson's sub-realtime soft-body simulation. By maintaining consistency across the robot, the task, and the teleoperation interface, we find that for both tasks the completion time in simulation and real are close, demonstrating that the simulation environment in OmniGibson is a good proxy for mobile manipulation in the real world, and that teleoperating with TeleMoMa provides a natural mechanism to collect demonstrations in sim.

2) Comparing Embodiments: We additionally compare how the completion time varies as we change the robot being teleoperated by maintaining the task, teleoperation interface and reality to be consistent. We compare Tiago and HSR on `re-shelve chips` and `open fridge` tasks and depict the results in Fig. 5(c, right). We observe that the higher number of degrees of freedom offered by Tiago compared to HSR allows more fluid motion during teleoperation and enables a more efficient (faster) completion of the task.

In simulation, we compare Tiago and Fetch on `cover table`, `slide chair`, and `pick pot` tasks and depict the results in Fig. 5(c, left). For the `pick pot` task, we enabled sticky grasping (creating a controllable constraint between hand and object) since the task would be infeasible otherwise for a single-armed robot like Fetch. We observe that Fetch is faster than Tiago on tasks requiring table-top manipulations, possibly due to Fetch's larger size and longer arms, making manipulation easier for users.

Hyperparameters	Value
Behavior Cloning (BC)	
train steps (x500)	500
batch size	32
optimizer	Adam
learning rate	1e-4
image & depth encoder	resnet-18
policy (w x d)	512x2
action parameterization	GMM
Recurrent BC (BC-RNN)	
train steps (x500)	500
batch size	16
optimizer	Adam
learning rate	1e-4
image & depth encoder	resnet-18
LSTM hidden dim	1000
LSTM num. layers	2
skill horizon	10
action parameterization	GMM

TABLE IV: Hyperparameters for the imitation policies (the hyperparameter values were kept consistent across tasks)

D. Imitation Learning Policy Hyperparameters

We performed imitation learning on one simulated (`slide chair` – Appendix C) and three real world tasks – `cover table`, `slide chair` and `serve bread`, that require synchronized hand and base motions. We used RoboMimic [3] for training the policies. Comprehensive details of the policy architecture and hyperparameters used for training are provided in Table IV. Note that the same hyperparameters were used across all tasks, and across simulation and real environments.