

TeleOpBench: A Simulator-Centric Benchmark for Dual-Arm Dexterous Teleoperation

Hangyu Li^{1,*} Qin Zhao^{1,2,*} Haoran Xu^{1,2} Xinyu Jiang¹ Qingwei Ben¹
Feiyu Jia¹ Haoyu Zhao¹ Liang Xu¹ Jia Zeng¹ Hanqing Wang¹ Bo Dai^{3,4}
Junting Dong^{1†} Jiangmiao Pang¹

¹Shanghai AI Laboratory ²Zhejiang University ³The University of Hong Kong ⁴Feeling AI
^{*}Equal Contribution [†]Project Leader

Abstract: Teleoperation is a cornerstone of embodied-robot learning, and bimanual dexterous teleoperation in particular provides rich demonstrations that are difficult to obtain with fully autonomous systems. While recent studies have proposed diverse hardware pipelines—ranging from inertial motion-capture gloves to exoskeletons and vision-based interfaces—there is still no unified benchmark that enables fair, reproducible comparison of these systems. In this paper, we introduce TeleOpBench, a simulator-centric benchmark tailored to bimanual dexterous teleoperation. TeleOpBench contains 30 high-fidelity task environments that span pick-and-place, tool use, and collaborative manipulation, covering a broad spectrum of kinematic and force-interaction difficulty. Within this benchmark we implement four representative teleoperation modalities—(i) MoCap, (ii) VR device, (iii) arm–hand exoskeletons, and (iv) monocular vision tracking—and evaluate them with a common protocol and metric suite. To validate that performance in simulation is predictive of real-world behavior, we conduct mirrored experiments on a physical dual-arm platform equipped with two 6-DoF dexterous hands. Across 10 held-out tasks we observe a strong correlation between simulator and hardware performance, confirming the external validity of TeleOpBench. TeleOpBench establishes a common yardstick for teleoperation research and provides an extensible platform for future algorithmic and hardware innovation.

Keywords: Teleoperation benchmark, Dual-arm dexterous teleoperation

1 Introduction

Recent breakthroughs in robot learning [1, 2, 3, 4] have been fueled by ever-growing repositories of human-demonstration data [5, 6, 7, 8, 9, 10, 11], which supply rich priors and reduce the sample complexity of learning in the real world. Teleoperation stands out as a pivotal data-acquisition paradigm, yielding precise yet natural manipulation trajectories that are indispensable for training high-fidelity control policies. Unlike single-arm grippers, humanoid dual-arm dexterous manipulators unlock the execution of intricate, fine-grained tasks, but their heightened kinematic complexity and need for tightly coordinated bimanual motion render teleoperation more challenging.

Recent advances in dual-arm dexterous teleoperation [12, 13] have showcased remarkable progress, leveraging a diverse suite of operator interfaces—from inertial and optical motion-capture setups [14, 15] to VR controllers [16, 17], upper-body exoskeletons [18, 19, 20], and purely vision-based trackers [21, 22]. Yet, despite these impressive demonstrations, the community still lacks a standardized benchmark that would enable rigorous, fair, and comprehensive comparisons across competing approaches. Because each system is tightly coupled to its own mix of teleoperation hardware, robot platform, and task environment, cross-method evaluation remains muddled, obscuring objective assessments of performance and ultimately hampering progress in dexterous teleoperation research.

To solve this, we introduce TeleOpBench, a novel simulator-based benchmark expressly designed for fair evaluation of dual-arm teleoperation systems. Because the simulator fixes both the robot

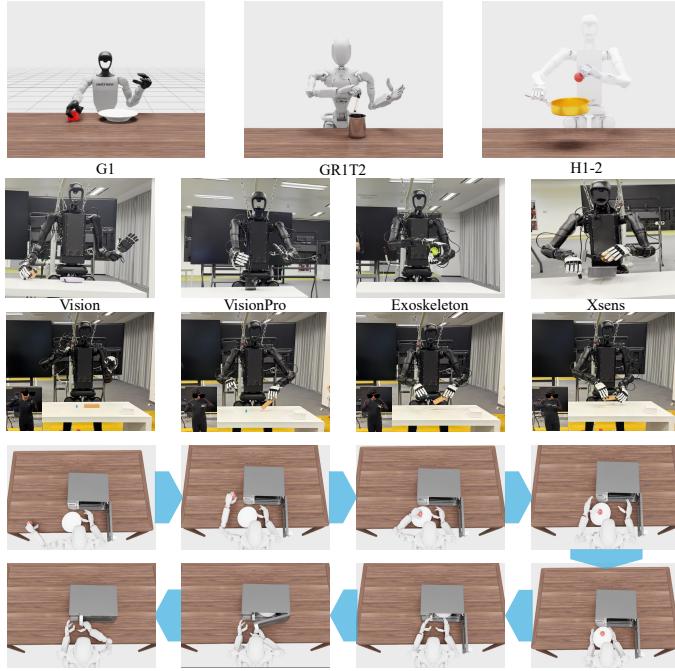


Figure 1: We present TeleOpBench, a simulation-based benchmark for bimanual dexterous teleoperation, and evaluate four representative teleoperation modalities across multiple robot platforms (row 1). Real-robot experiments (row 2) demonstrate four teleoperation capabilities. Our teleoperation pipelines support fine-precision manipulation in the real world—for example, the left hand grasps a block while the right hand simultaneously inserts a smaller block (row 3)—and can execute long-horizon sequences, such as retrieving a tomato-laden plate from a microwave with the right hand and transferring the tomatoes to a table with the left (rows 4 and 5).

morphology and the task environment, it eliminates the hardware and scene variability that plagues real-robot comparisons, making it uniquely suited for systematic assessment. TeleOpBench couples a broad spectrum of task environments with multiple teleoperation interfaces within a single, coherent framework. Concretely, we provide 30 progressively harder tasks (ranging from simple cube pick-and-place to long-horizon routines such as lifting a pot lid and transferring fruit from the pot to an external dish) and the task suite can be easily extended or customized by users. In addition, under a unified, modular interface, we implement four representative dual-arm teleoperation modalities: inertial motion capture, VR controllers, upper-body exoskeletons, and vision-only. Researchers can effortlessly plug new teleoperation pipelines into this framework and benchmark them under exactly the same conditions, enabling truly fair comparisons.

Leveraging TeleOpBench, we conduct a comprehensive evaluation of the four teleoperation modalities, reporting task-wise success rates and completion times across diverse tasks. We further replicate nearly identical scenarios on a physical dual-arm platform and gather real-world performance metrics for each teleoperation system. The strong correlation between simulator and hardware results confirms that TeleOpBench faithfully predicts real-robot outcomes, underscoring its value as a rigorous benchmarking tool. All code and task assets will be released open-source to foster transparent, reproducible research and to accelerate progress in dexterous teleoperation.

In summary, this paper makes the following contributions:

1. We introduce a dedicated benchmark, TeleOpBench, for dual-arm dexterous teleoperation, enabling rigorous, fair, and comprehensive comparisons across competing systems.
2. We implement four representative teleoperation pipelines—motion-capture, VR controllers, upper-body exoskeletons, and vision-only within a single modular framework.

3. Extensive experiments on both TeleOpBench and a real dual-arm platform reveal a strong correlation between simulated and physical performance, substantiating the benchmark’s fidelity and practical value.

2 Related Work

Bimanual dexterous teleoperation. Teleoperation emerges as a crucial paradigm for acquiring robot operation data. Current teleoperation systems have evolved from grippers [23, 24, 25, 26], single-arm setups [27, 28, 11] to bimanual dexterous hands [12, 13]. Compared with gripper-based or single-arm set-ups, bimanual, multi-finger platforms unlock far more intricate manipulation skills, yet they also amplify the difficulty of teleoperation. To meet these new demands, researchers have recently explored a spectrum of input modalities—notably exoskeletons, motion capture (MoCap), virtual-reality (VR) devices, and purely vision-based interfaces. Exoskeleton-driven teleoperation [18, 19, 29, 23, 25, 30, 26, 20] removes the need for a kinematically identical master robot; joint-level matching or inverse-kinematics (IK) mapping allows operators to drive the robot with high precision and low latency. When paired with motion-sensing gloves, these systems can render truly dexterous manipulation. VR-based approaches [31, 32, 33, 16, 17] employ hand-held controllers or egocentric cameras to recover wrist pose and finger keypoints, which are then transformed via IK into robot joint targets, offering a cost-effective yet immersive control loop. MoCap systems [34, 35, 36, 37, 38, 39, 14, 15, 40]—whether inertial or optical—track full arm-hand kinematics at high frequency, achieving both high accuracy and bandwidth, but at the expense of specialized hardware and calibration effort. Vision-only methods [21, 22] estimate wrist and finger pose directly from monocular cameras [41, 42]; although they currently lag behind MoCap in precision and update rate, they dramatically reduce deployment cost and complexity, making teleoperation more accessible. Together, these modalities chart a rich design space for bimanual teleoperation, each balancing fidelity, latency, and affordability in different ways—trade-offs that our benchmark seeks to evaluate systematically.

Robotics benchmark. A well-designed benchmark provides a standardized, reproducible and equitable environment for assessing different approaches, which substantially promotes the development of the field [43, 44, 45, 46, 47]. For robotics-related tasks, real-world experiments introduce significant uncertainty from hardware setups, lighting conditions and evaluation task configurations. Thus, many studies have developed simulation benchmarks as alternatives [48, 49, 50, 51, 52, 53, 54, 52, 55, 56, 57]. In particular, a growing number of simulation-based evaluation platforms have emerged for robotic reinforcement learning [58, 59, 60, 61, 50, 62, 53, 63, 64] and imitation learning [65, 66, 67, 68]. Comprehensive evaluation of teleoperation systems aims to quantify the performance, reliability and usability of the human operators controlling robotic platforms through various interfaces. Existing research on evaluating teleoperation systems has explored a diverse range of robotic platforms, input interfaces, task environments and realities (*i.e.*, real or simulation), which makes the cross-method fair comparison and reproducibility infeasible. Inspired by previous robotic benchmarks, we propose a simulation-centric evaluation platform named TeleOpBench for bimanual dexterous teleoperation benchmarking, which supports various input interfaces, robot entities, and a wide range of customizable tasks.

3 TeleOpBench

We introduce TeleOpBench, a simulator-based benchmark purpose-built for impartial evaluation of dual-arm teleoperation systems. Figure 2 shows an overview of TeleOpBench. Leveraging the simulator’s controllability, TeleOpBench provides 30 task environments spanning a wide difficulty spectrum—from elementary cube pick-and-place to long-horizon routines such as lifting a pot lid and transferring fruit from the pot to an external dish (Section 3.1). Furthermore, four representative teleoperation pipelines—inertial motion capture, VR controllers, upper-body exoskeletons, and vision-only hand tracking—are implemented under a unified, modular interface and instantiated on three different dual-arm robots (Section 3.2).

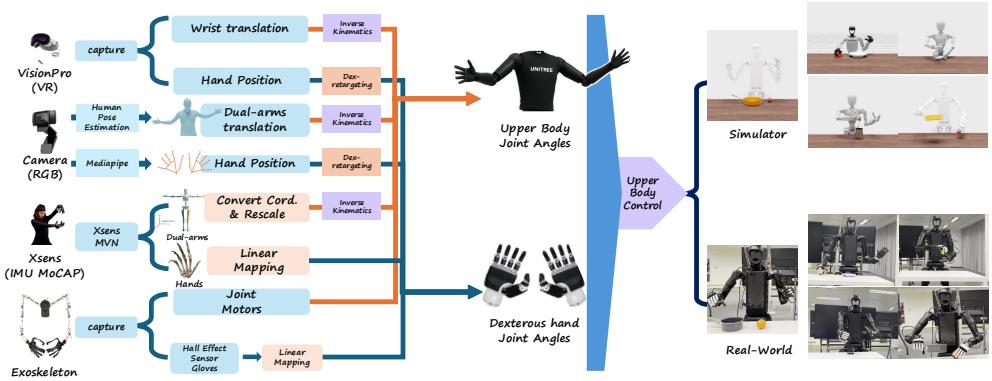


Figure 2: The overview of the proposed TeleOpBench, where we unify four operator interfaces in both simulation and real-world realities for dual-arm dexterous teleoperation.

3.1 Task environments

We employ NVIDIA Isaac Sim as our simulation platform because its high-performance PhysX engine and photorealistic renderer enable the construction of environments that closely approximate real-world conditions. Each scene features a humanoid robot fitted with bimanual dexterous hands and the task-relevant objects; operators are instructed to execute the required manipulations exactly as specified. For every trial, we record both task success and completion time, which together constitute our primary performance metrics.

Humanoid robots. For a comprehensive hardware evaluation, we employ three commercially available humanoid platforms—Unitree H1-2, Fourier GR1-T2, and Unitree G1. Unitree H1-2 is a full-size humanoid with 7 DoF in each arm. The base model is equipped with Inspire Dexterous Hands with five articulated fingers and 6 DoF for fine manipulation. Fourier GR1-T2 matches2 in overall stature and arm kinematics (7 DoF)DoF),t ships with Fourier’s native dexterous hands, which provide five fingers and 6 DoF. Unitree G1 adopts a far more compact form factor and is equipped with lightweight three-finger hands offering 4 DoF. The trio spans a meaningful range of scale and hand design, enabling a systematic assessment of teleoperation across diverse humanoid robots.

Task setting. TeleOpBench offers a comprehensive, multi-tiered suite of 30 bimanual dexterous-manipulation tasks. The tasks are hierarchically organized by complexity—e.g., whether they require coordinated two-hand interaction or long-horizon sequencing—so that both coarse- and fine-grained teleoperation modalities can be evaluated in an appropriately graduated setting. This progression is essential: if all tasks were uniformly difficult, lower-precision interfaces such as monocular-vision tracking would fail wholesale; if all tasks were trivial, higher-precision methods such as inertial MoCap would be indistinguishable from less capable alternatives.

To facilitate customization and encourage community contributions, we provide a modular, extensible asset library with flexible APIs that let researchers instantiate new scenarios or tune task difficulty with minimal effort. Every task includes explicit success criteria, ensuring that results are assessed under consistent, objective standards. In simulation, we replicate real-world physics as closely as possible to match object masses, friction coefficients, and other dynamics so that each scenario remains physically and pragmatically faithful. A complete list of tasks and their configurations is shown in Table 1.

Observation. Our simulation framework records a rich set of observations for every teleoperation episode, enabling downstream imitation-learning studies and further amplifying the value of TeleOpBench. The logged data include (a) robot-state vectors—joint positions, angles, velocities, and related kinematics; (b) camera streams—RGB images from a head-mounted, first-person camera and a

Task Name	Completion Criteria	Task Name	Completion Criteria
push_cube	push the cube into the left blue area	pot_bimanual	lift the pot with both hands
pick_cube	pick up and place the cube down	pot_tomato	lift the pot and add tomato into it
pick_place_cube	pick up the cube and place it on the plate	pot_tray	place the pot onto the tray with both hands
uprear_cup	pick up the cup and place it upright	stack_boxes	lift the box and stack it onto another box
ball_trashcan	pick up the ball and place it into the trashcan	pan_hearth	open the lid and place the pan on the hearth
rotate_faucet	rotate the faucet 90 degrees	tidyup_table	place objects into the basket in order
rotate_hearth	press and turn the hearth knob	pour_water	pour water from the kettle into the cup
open_microwave	pull open the microwave door	pot_tomato_out	open the pot and remove the tomato from it
close_microwave	close the microwave door	plate_oven	open the oven, place the plate, close it
open_drawer	pull open the drawer	pot_tomato_plate	open the lid and place the tomato on the plate
close_drawer	close the drawer	pen_brushpot	pick up the pen and place it into the container
lift_mug	lift the lid of the mug	drawer_book	open the drawer and place the books inside
open_laptop	open the laptop	bread_toaster	place the bread into the toaster and press the button
ball_mug	pick up the ball and place it into the mug	stack_toyblocks	assemble the toy blocks in sequence
ball_bimanual	pass the ball from one hand to another	twist_bottle_cap	pick up the bottle and twist the cap on

Table 1: The list of task names and completion criteria of TeleOpBench.

fixed third-person camera facing the workbench; and (c) task-level environment metadata—precise object positions and orientations.

3.2 Modular Teleoperation interface

We implement four representative teleoperation pipelines—monocular vision, MoCap, VR, and exoskeleton—under a unified, modular interface.

3.2.1 Vision-based

Unlike prior vision-based teleoperation methods, we decouple arm-and-wrist pose estimation from hand-keypoint estimation, resulting in a more robust and higher-precision visual interface. Concretely, our vision pipeline consists of three core modules: human body parameter estimation and scale, upper-body limb motion control, and hand control.

Human-body parameter estimation and scaling. To reduce the sensitivity of teleoperation accuracy to inter-subject anthropometric differences, we follow the philosophy of PHC [69] and build a keypoint-constrained parameter-scaling model. In contrast to PHC, we solve the scaling parameters only once under a neutral T-pose, which is sufficient for subsequent sessions of the same operator. We select four anatomically consistent landmarks—pelvis, shoulders, wrists, and head—on both the human operator and the robot. The body parameter is obtained by minimizing the keypoint alignment error, formulated as:

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^{10}} \sum_l \|f_{\text{SMPL}}^l(\beta, \theta_0) - R^l\|_2^2, \quad (1)$$

where $f_{\text{SMPL}}^l(\beta, \theta)$ denotes the 3D position of the l -th human landmark generated by the parametric human model SMPL [70] from shape parameter β and pose parameter θ . The vector θ_0 represents the T-pose and R^l denotes the 3D position of the l -th landmark on the robot. We employ gradient-based optimization to obtain the optimal body parameter β^* . Under β^* , we calculate the scaling factors between each human joint link and its corresponding robotic counterpart, thereby deriving a set of optimal scale parameters s^* .

Upper-body limb motion control. We utilize SMPLer-X [71] to capture the teleoperator’s SMPL pose parameter, then compute joint positions under the optimized body parameter β^* . These positions are scaled by the derived factors s^* to obtain robotic translations, followed by PINK [72] based inverse kinematics solving to derive all DoF except finger joints.

Hand control. We provide two control schemes. Scheme 1: Directly use the finger rotations from SMPLer-X to calculate the corresponding Euler angles, then set the robotic hand’s DoF values based on these angles. Scheme 2: Utilize MediaPipe to capture finger keypoint positions, then apply vector optimizers via Dex-Retargeting [22]—a highly versatile and computationally efficient motion retargeting library—achieving significantly improved performance. We adopt Scheme 2 throughout

all experiments. Finally, we implement a Kalman filter to smooth the robot’s DoF, significantly reducing jitter-related instability in motion execution.

3.2.2 MoCap-based

Hardware. We employ the Xsens MVN system [73] as our motion capture solution. For capturing limb movements, the core sensor assembly includes seventeen IMUs strategically attached to corresponding human body segments. Furthermore, Xsens Metagloves by Manus are utilized to precisely capture intricate hand motions through a sophisticated hand model, providing 20 degrees of freedom (DOFs) for each hand individually. Upon wearing the Xsens suit and completing the calibration process, data capturing the positions and orientations of the seventeen body segments, as well as detailed finger joint angles, is directly accessible through the MVN system.

Arm control. The raw data obtained from the Xsens system is initially defined within its proprietary global coordinate system. Thus, the first necessary step involves transforming this data into the robot’s coordinate system. Within our teleoperation setup, the robot’s lower body is immobilized; hence, we define the robot’s coordinate reference frame at the pelvis joint, with the robot’s forward-facing direction aligned with the positive X-axis and the vertical direction aligned with the Z-axis, following a right-handed coordinate system convention. To address the inherent skeletal differences between the humanoid robot and the human operator—which could lead to significant discrepancies in motion if directly used in IK—a joint-specific rescaling method is introduced. This rescaling approach calculates and updates scaling parameters in real-time upon receiving the initial raw data. Subsequently, it adjusts joint lengths individually, converting the upper limb and arm joint coordinates accurately into the robot’s coordinate frame. Finally, we compute the robot’s joint poses using Closed-Loop Inverse Kinematics (CLIK), ensuring precise and robust teleoperated control.

Hand control. The Manus glove provides detailed axis-angle data representing 20 degrees of freedom per hand. This data encompasses the flexion and abduction/adduction movements of the metacarpophalangeal (MCP) joints (connecting each finger to the palm), as well as the flexion of the proximal interphalangeal (PIP) joints (the intermediate joints of each finger) and the distal interphalangeal (DIP) joints (nearest to each fingertip). The captured finger motion data from the Manus gloves is subsequently mapped to the joint angle constraints defined by the robotic dexterous hand (dexhand), ensuring accurate and precise finger control.

3.2.3 VR-based

The VR-based teleoperation system consists of upper-body limb motion control and hand control.

Upper-body limb motion control. We utilize the Apple VisionPro for hand, wrist, and head tracking. The tracking adheres to the OpenXR coordinate system. The wrist and head poses are first transformed into the robot’s coordinate frame, and the wrist offset relative to the head is subsequently converted into an offset relative to the pelvis. We exclusively feed the wrist translation data to the IK algorithm based on Pink [74], which computes all degrees of freedom except finger joints.

Hand control. To enhance manual dexterity across different teleoperators, we measure the distal phalanx lengths of each operator’s fingers and scale them proportionally to match the corresponding robotic finger segments, resulting in a scaling factor $s^* \in \mathbb{R}^5$. Subsequently, following Open-Television [75], we employ vector-based optimizers to generate robot-hand joint commands via the dexterous-retargeting framework of AnyTeleop [22].

3.2.4 Exoskeleton-based

We propose a framework for designing isomorphic exoskeleton systems tailored to diverse humanoid platforms, enabling high-precision teleoperation across simulated and physical environments. Building on principles from HOMIE [20], each exoskeleton is customized to replicate the kinematic structure of its target humanoid’s upper body, utilizing servo-driven joints to synchronize human operator movements with robotic counterparts in real time. Integrated motion-sensing gloves equipped

with Hall-effect sensors provide 15 degrees of freedom (DoF) per-hand tracking for dexterous manipulation. By directly mapping operator kinematics to the humanoid’s joints—bypassing inverse kinematics (IK) approximations—our platform-specific exoskeletons eliminate algorithmic errors while enhancing operational bandwidth and positional accuracy.

4 Experiments



Figure 3: From top to bottom, we illustrate the four teleoperation modalities executing the following tasks: ball_trashcan, pen_brushpot, ball_bimanual, and pot_bimanual.

In this section, we evaluate the effectiveness of TeleOpBench. First, we showcase the performance of four teleoperation schemes across both simulated and real-world tasks. We then carry out a systematic comparison in the simulator and analogous task settings on physical hardware.

Qualitative results. Figure 3 presents qualitative results of the four teleoperation modalities across both simulated and real-world task settings. Our vision-based system achieves a high success rate in grasping tasks, e.g., accurately placing a ball into a trash bin. The VR-based system successfully places a slender pen into a pen holder. The exoskeleton-based system allows the right hand to pick up a ball and transfer it to the left hand for a stable grasp. The Xsens motion-capture system supports high-precision manipulation—e.g., picking up a slender pen into a pen holder.

Task	Vision-Based		VR-Based		Exoskeleton		Xsens	
	Succ (%)	Time (s)	Succ (%)	Time (s)	Succ (%)	Time (s)	Succ (%)	Time (s)
1	80	13.64	100	15.32	90	16.42	100	6.32
2	100	34.66	100	15.54	100	12.69	100	7.34
3	80	33.00	100	12.46	90	20.28	100	7.87
4	40	56.50	80	21.67	80	16.91	90	11.19
5	70	35.96	100	12.51	100	15.48	100	14.52
6	60	52.75	100	15.13	90	17.86	100	9.97
7	0	—	0	—	80	21.06	100	14.70
8	50	36.64	100	9.62	100	7.85	90	16.36
9	10	24.87	90	41.72	80	37.86	100	11.38
10	0	—	70	57.32	80	23.92	100	12.63

Table 2: Performance comparison of teleoperation systems across tasks in simulation.

Quantitative results in simulation tasks. We select ten representative tasks of varying difficulty from the TeleOpBench: (1) push_cube, (2) pitch_cube, (3) pick_place_cube, (4) uprear_cup, (5) ball_trashcan, (6) ball_mug, (7) ball_bimanual, (8) pot_bimanual, (9) pot_tomato_plate, and (10) pen_brushpot. Full task descriptions refer to Table 1. A user study involving four participants was conducted; Task-level success rates and completion times are summarized quantitatively in Table 2.

For vision-based methods, monocular camera keeps the setup simple, but low frame-rate, coarse wrist-orientation estimates, and occlusion limit it to easy tasks, yielding the largest completion times and poor success on Tasks 4 and 7. For VR, accurate wrist/hand tracking gives strong grasps, yet Task 7 fails entirely because hand-over-hand occlusion breaks pose estimation. The exoskeleton-based method, with a kinematically aligned design and direct DoF mapping to the robot, delivers smooth control and performs well in most tasks. However, due to limited capability in lateral elbow movement, it shows increased time consumption in Task 1 Push the cube. The Xsens-based method excels in both smoothness and motion precision. It completes the tasks accurately and typically with the least time cost. However, it is also the most expensive among the four teleoperation systems.

Task	Vision-Based		VR-Based		Exoskeleton		Xsens	
	Succ (%)	Time (s)	Succ (%)	Time (s)	Succ (%)	Time (s)	Succ (%)	Time (s)
1	100	14.41	100	15.29	90	18.47	100	10.44
2	70	30.79	100	9.82	100	9.77	100	8.31
3	80	14.95	100	10.16	100	8.48	100	6.12
4	40	24.79	70	14.32	80	15.22	90	11.33
5	60	23.11	100	13.57	80	12.43	100	6.91
6	20	26.21	90	13.86	60	16.75	100	8.18
7	0	—	0	—	70	24.82	90	12.90
8	40	26.34	100	11.18	100	5.49	100	12.02
9	10	53.34	90	36.32	80	22.43	100	17.97
10	0	—	80	24.31	70	27.49	100	16.47

Table 3: Performance comparison of teleoperation systems across tasks in real world.

Quantitative results in physical tasks. We reproduce the task suite on physical robots and evaluate all four teleoperation pipelines with the identical metric suite; the resulting quantitative scores are summarized in Table 3. Figure 4 presents completion-time curves for simulation and real world. Note that Tasks with one teleoperation success rate below 20% are excluded from the plotted curves to ensure the reliability of the curves. The two domains exhibit a strong positive correlation: the vision-tracking interface consistently requires the longest execution time (blue curve), the inertial-MoCap pipeline is the fastest (red curve), and the VR and exoskeleton interfaces cluster in between. This close alignment between simulated and real-world performance confirms that TeleOpBench reliably predicts practical outcomes and therefore offers substantial utility to the community.

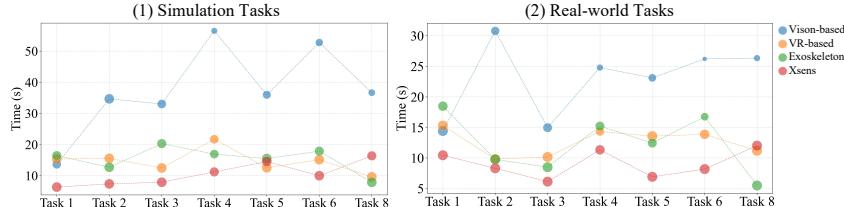


Figure 4: Completion-time and success rate curves for simulation and real world. The size of each circle reflects the corresponding task success rate.

5 Conclusion

We present TeleOpBench, a simulator-centric benchmark for bimanual dexterous teleoperation, providing a fair, reproducible platform for cross-system comparison. TeleOpBench contains 30 high-fidelity task environments spanning a broad spectrum of difficulty. Within this suite, we implement four representative teleoperation modalities in a unified, modular framework. Extensive experiments in both simulation and on physical hardware reveal a strong correlation between simulated and real-world performance, validating the benchmark’s external fidelity and underscoring its practical value for future research.

6 Limitations

TeleOpBench presently targets upper-body teleoperation in predominantly tabletop settings. A natural next step is to build a loco-manipulation benchmark that couples dexterous arm-hand control with lower-body locomotion, thereby testing whole-body teleoperation pipelines. A second promising direction is to incorporate haptic-feedback interfaces. All modalities evaluated in this study lack tactile feedback; adding haptic would enable the assessment of finer force-controlled tasks and further broaden the benchmark’s applicability.

Acknowledgments

If a paper is accepted, the final camera-ready version will (and probably should) include acknowledgments. All acknowledgments go at the end of the paper, including thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

- [1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [2] 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.
- [3] 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.
- [4] H. Zhao, H. Wang, X. Zhao, H. Wang, Z. Wu, C. Long, and H. Zou. Automated 3d physical simulation of open-world scene with gaussian splatting. *arXiv preprint arXiv:2411.12789*, 2024.
- [5] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.
- [6] H. Xiong, R. Mendonca, K. Shaw, and D. Pathak. Adaptive mobile manipulation for articulated objects in the open world. *arXiv preprint arXiv:2401.14403*, 2024.
- [7] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.
- [8] K. Shaw, S. Bahl, and D. Pathak. Videodex: Learning dexterity from internet videos. *CoRL*, 2022.
- [9] Y. Zhu, A. Lim, P. Stone, and Y. Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024.
- [10] R. Mendonca, S. Bahl, and D. Pathak. Structured world models from human videos. *RSS*, 2023.
- [11] J. Wang, Y. Qin, K. Kuang, Y. Korkmaz, A. Gurumoorthy, H. Su, and X. Wang. CyberDemo: Augmenting Simulated Human Demonstration for Real-World Dexterous Manipulation. *arXiv preprint arXiv: 2312.09237*, 2024.
- [12] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik. Learning visuotactile skills with two multifingered hands. *arXiv:2404.16823*, 2024.

- [13] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto. Open teach: A versatile teleoperation system for robotic manipulation, 2024.
- [14] S. Dafarra, K. Darvish, R. Grieco, G. Milani, U. Pattacini, L. Rapetti, G. Romualdi, M. Salvi, A. Scalzo, I. Sorrentino, et al. icub3 avatar system. *arXiv preprint arXiv:2203.06972*, 2022.
- [15] R. Cisneros, M. Benallegue, K. Kaneko, H. Kaminaga, G. Caron, A. Tanguy, R. Singh, L. Sun, A. Dallard, C. Fournier, et al. Team janus humanoid avatar: A cybernetic avatar to embody human telepresence. In *Toward Robot Avatars: Perspectives on the ANA Avatar XPRIZE Competition, RSS Workshop*, 2022.
- [16] J. Chagas Vaz, D. Wallace, and P. Y. Oh. Humanoid loco-manipulation of pushed carts utilizing virtual reality teleoperation. In *ASME International Mechanical Engineering Congress and Exposition*, 2021.
- [17] S. P. Arunachalam, I. Güzey, S. Chintala, and L. Pinto. Holo-dex: Teaching dexterity with immersive mixed reality. *arXiv preprint arXiv:2210.06463*, 2022.
- [18] J. Ramos and S. Kim. Humanoid dynamic synchronization through whole-body bilateral feedback teleoperation. *IEEE Transactions on Robotics*, 2018.
- [19] Y. Ishiguro, T. Makabe, Y. Nagamatsu, Y. Kojio, K. Kojima, F. Sugai, Y. Kakiuchi, K. Okada, and M. Inaba. Bilateral humanoid teleoperation system using whole-body exoskeleton cockpit tablis. *IEEE Robotics and Automation Letters*, 2020.
- [20] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.
- [21] 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)*, pages 9164–9170. IEEE, 2020.
- [22] 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. In *Robotics: Science and Systems*, 2023.
- [23] 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.
- [24] 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)*, pages 1–8. IEEE, 2023.
- [25] 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.
- [26] 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.
- [27] S. P. Arunachalam, I. Güzey, S. Chintala, and L. Pinto. Holo-dex: Teaching dexterity with immersive mixed reality. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5962–5969. IEEE, 2023.
- [28] Y. Qin, H. Su, and X. Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7(4):10873–10881, 2022.

- [29] M. Schwarz, C. Lenz, A. Rochow, M. Schreiber, and S. Behnke. Nimbro avatar: Interactive immersive telepresence with force-feedback telemanipulation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5312–5319, 2021.
- [30] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu. Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15031–15038. IEEE, 2024.
- [31] J. I. Lipton, A. J. Fay, and D. Rus. Baxter’s homunculus: Virtual reality spaces for teleoperation in manufacturing. *IEEE Robotics and Automation Letters*, 3(1):179–186, 2017.
- [32] E. Rosen, D. Whitney, M. Fishman, D. Ullman, and S. Tellex. Mixed reality as a bidirectional communication interface for human-robot interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11431–11438. IEEE, 2020.
- [33] P. Ponomareva, D. Trinitatova, A. Fedoseev, I. Kalinov, and D. Tsetserukou. Grasplook: a vr-based telemانipulation system with r-cnn-driven augmentation of virtual environment. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 166–171. IEEE, 2021.
- [34] W. Zhao, J. Chai, and Y.-Q. Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *Proceedings of the ACM SIGGRAPH/eurographics symposium on computer animation*, pages 33–42, 2012.
- [35] A. D. Dragan, K. C. Lee, and S. S. Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013.
- [36] H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, and S.-C. Zhu. A glove-based system for studying hand-object manipulation via joint pose and force sensing. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6617–6624. IEEE, 2017.
- [37] K. Darvish, Y. Tirupachuri, G. Romualdi, L. Rapetti, D. Ferigo, F. J. A. Chavez, and D. Pucci. Whole-body geometric retargeting for humanoid robots. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, 2019.
- [38] H. Liu, Z. Zhang, X. Xie, Y. Zhu, Y. Liu, Y. Wang, and S.-C. Zhu. High-fidelity grasping in virtual reality using a glove-based system. In *2019 international conference on robotics and automation (icra)*, pages 5180–5186. IEEE, 2019.
- [39] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021.
- [40] M. Mosbach, K. Moraw, and S. Behnke. Accelerating interactive human-like manipulation learning with GPU-based simulation and high-quality demonstrations. In *Humanoids*, 2022.
- [41] H. Zhao, H. Wang, C. Yang, and W. Shen. Chase: 3d-consistent human avatars with sparse inputs via gaussian splatting and contrastive learning. *arXiv preprint arXiv:2408.09663*, 2024.
- [42] H. Zhao, C. Yang, H. Wang, X. Zhao, and W. Shen. Sg-gs: Photo-realistic animatable human avatars with semantically-guided gaussian splatting. *arXiv preprint arXiv:2408.09665*, 2024.
- [43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

- [45] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- [46] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [47] Q. Vuong, S. Levine, H. R. Walke, K. Pertsch, A. Singh, R. Doshi, C. Xu, J. Luo, L. Tan, D. Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.
- [48] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, V. Kumar, and W. Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.
- [49] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. P. Lillicrap, and M. A. Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.
- [50] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *arXiv preprint arXiv:1910.10897*, 2019.
- [51] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.
- [52] V. Caggiano, H. Wang, G. Durandau, M. Sartori, and V. Kumar. Myosuite—a contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*, 2022.
- [53] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.
- [54] S. Dasari, J. Wang, J. Hong, S. Bahl, Y. Lin, A. Wang, A. Thankaraj, K. Chahal, B. Calli, S. Gupta, et al. Rb2: Robotic manipulation benchmarking with a twist. *arXiv preprint arXiv:2203.08098*, 2022.
- [55] F. Al-Hafez, G. Zhao, J. Peters, and D. Tateo. Locomujoco: A comprehensive imitation learning benchmark for locomotion. *6th Robot Learning Workshop at NeurIPS*, 2023.
- [56] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *The International Journal of Robotics Research*, page 02783649241304789, 2023.
- [57] C. Sferrazza, D.-M. Huang, X. Lin, Y. Lee, and P. Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506*, 2024.
- [58] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI Gym, 2016.
- [59] S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, N. Heess, and Y. Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. ISSN 2665-9638. doi:<https://doi.org/10.1016/j.simpa.2020.100022>. URL <https://www.sciencedirect.com/science/article/pii/S2665963820300099>.
- [60] Y. Tassa, S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, and N. Heess. dm_control: Software and tasks for continuous control, 2020.

- [61] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv:1606.01540*, 2016.
- [62] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.
- [63] Y. Lee, E. S. Hu, and J. J. Lim. IKEA furniture assembly environment for long-horizon complex manipulation tasks. In *ICRA*, 2021. URL <https://clvrai.com/furniture>.
- [64] V. Caggiano, H. Wang, G. Durandau, M. Sartori, and V. Kumar. Myosuite—a contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*, 2022.
- [65] R. Memmesheimer, I. Kramer, V. Seib, and D. Paulus. Simitate: A hybrid imitation learning benchmark. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5243–5249. IEEE, 2019.
- [66] F. Al-Hafez, G. Zhao, J. Peters, and D. Tateo. Locomujoco: A comprehensive imitation learning benchmark for locomotion. *arXiv preprint arXiv:2311.02496*, 2023.
- [67] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.
- [68] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.
- [69] Z. Luo, J. Cao, K. Kitani, W. Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023.
- [70] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.
- [71] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang, C. C. Loy, L. Yang, and Z. Liu. SMPLER-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2023.
- [72] S. Caron, Y. De Mont-Marin, R. Budhiraja, S. H. Bang, I. Domrachev, and S. Nedelchev. Pink: Python inverse kinematics based on Pinocchio, 2025. URL <https://github.com/stephane-caron/pink>.
- [73] D. Roetenberg, H. Luinge, P. Slycke, et al. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep*, 1(2009):1–7, 2009.
- [74] S. Caron, Y. De Mont-Marin, R. Budhiraja, S. H. Bang, I. Domrachev, and S. Nedelchev. Pink: Python inverse kinematics based on Pinocchio, 2024. URL <https://github.com/stephane-caron/pink>.
- [75] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.