DexDirect: Direct Kinesthetic Arm Guidance for Efficient Dexterous Demonstration Collection

Beom Jun Kim[1], Shiu-Jen Wang[1], Jonathan Liu[2], Alvin Zhu[2,3], Quanyou Wang[1], Hanzhang Fang[2,3], Feng Xu[1], Mingzhang Zhu[1], Yuchen Cui[3†], and Dennis W. Hong[1†]

***Abstract*— Scalable collection of dexterous manipulation demonstrations remains a major bottleneck for robot learning. High-fidelity interfaces often require costly hardware and extensive setup, while low-setup, low-cost alternatives tend to provide less precise control and impose greater cognitive workload on operators. We present DexDirect, a direct kinesthetic arm guidance for efficient dexterous demonstration collection. The operator drags a 6-DoF gravity-compensated robot arm directly by a handle, while a single webcam retargets operator's other hand onto a 16 joints 13-DoF dexterous robot hand. User studies suggest DexDirect collects 17.2× and 3.2× more successful demonstrations compared to purely vision (AnyTeleop) and pose-tracking (TeleDex) baselines. An adapted NASA-TLX shows DexDirect greatly reduces mental demand, effort, and frustration, despite raising physical demand. A diffusion policy trained on DexDirect demonstrations reaches a 90% success rate on a cube pick-and-place task. These results suggest that direct kinesthetic arm guidance combined with vision-based hand retargeting provides an efficient low-setup and scalable interface for collecting dexterous manipulation demonstrations.**

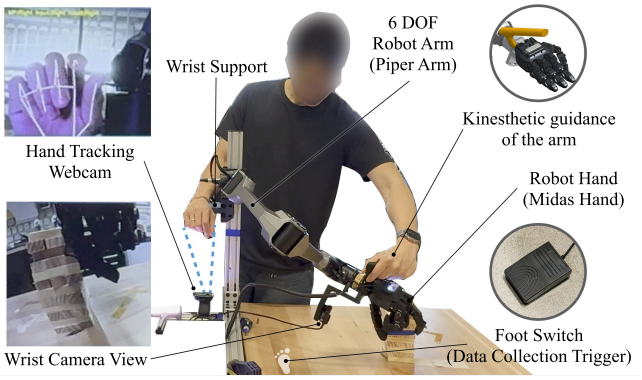


Fig. 1: DexDirect combines direct kinesthetic guidance of the robot arm with vision-based retargeting of the operator's free hand. One hand guides the robot through a wrist-mounted handle, while the other commands the dexterous hand.

I. Introduction

Learning robot manipulation policies from demonstrations has become a central paradigm for acquiring complex, contact-rich behaviors. The performance of these policies depends critically on the scale and quality of the collected demonstrations: an effective data-collection interface should produce accurate, robot-executable trajectories while remaining fast, inexpensive, and easy to deploy. In practice, however, it is difficult to achieve all of these objectives simultaneously. Interfaces that provide precise control and rich interaction feedback typically require specialized hardware and extensive setup, whereas lightweight alternatives often sacrifice control fidelity or increase the burden on the operator.

For parallel-jaw grippers, this trade-off has been mitigated by mature and scalable interfaces. Lightweight handheld devices allow demonstrations to be collected in diverse environments [1], while isomorphic leader arms provide accurate and intuitive joint-space control [2, 3]. Extending these solutions to multi-fingered hands is considerably more difficult. Although anthropomorphic robot hands resemble human hands, they differ in link geometry, joint axes, ranges of motion, compliance, and actuation. Moreover, the operator must simultaneously command the global 6-DoF motion of the arm and the high-dimensional articulation of the fingers, while regulating contacts that are often perceived only through visual feedback. These factors make dexterous demonstration collection slow, operator-dependent, and difficult to scale [4, 5].

Existing interfaces address this bottleneck by trading setup cost for demonstration fidelity. At one end are hardware-based interfaces, such as exoskeletons, that instrument the operator's hand and retarget measured joint angles to the robot [6]. Some additionally retain contact forces through using linkage mechanisms [7–9]. Although these systems differ mechanically, they all rely on dedicated linkage hardware: many require operator-specific fitting or calibration, while even operator-agnostic designs such as DexEXO [9] must be redesigned or refabricated for a new robot.

At the other end are vision-based and pose-tracking approaches that eliminate dedicated hardware altogether [5, 10, 11]. A commodity webcam or depth camera is sufficient to begin collecting demonstrations, while wearable pose trackers, such as a smartphone or headset, can improve tracking quality without introducing custom mechanical interfaces [12, 13]. We use this class of systems as the reference point for low-setup cost.

Low-setup systems, however, generally rely on a tracking-based control loop in which the operator moves a human

† denotes equal advising.

[1]Department of Mechanical and Aerospace Engineering, [2]Department of Electrical and Computer Engineering, [3]Department of Computer Science, [4]Department of Electrical and Computer Engineering, UCLA, Los Angeles, CA, USA.

hand or handheld device and the robot follows the estimated motion. Consequently, demonstration performance is affected by factors such as tracking latency, pose estimation error, workspace mismatch, and the cognitive burden of maintaining a coordinate mapping between the operator and the robot.

A natural way to avoid this proxy loop is to guide the robot directly. Kinesthetic teaching has been shown to provide an intuitive interface for non-expert users while recording configurations physically attained by the robot [14–16]. Applying the same approach to dexterous fingers, however, is non-trivial because manually guiding many small joints is mechanically cumbersome.

This asymmetry motivates our hybrid design. We present DexDirect, a low-setup interface for efficient dexterous demonstration collection that combines direct kinesthetic arm guidance with vision-based finger retargeting. The operator guides a gravity-compensated 6-DoF robot arm through a handle rigidly attached to the robot wrist, allowing the arm trajectory to be recorded directly from the robot joint encoders rather than inferred through external tracking. At the same time, a single webcam tracks the operator's other hand and retargets it to a 16 joints 13-DoF dexterous robot hand using the same class of commodity vision pipeline adopted by prior vision-based systems. As a result, DexDirect requires no leader arm, no operator-to-robot calibration for arm teleoperation, and no arm-side retargeting, while preserving the simple deployment of webcam-based systems.

In summary, our contributions are:

- **A low-setup, leader-less interface for dexterous demonstration collection.** DexDirect combines direct kinesthetic arm guidance with vision-based hand retargeting, eliminating separate leader hardware, arm-side tracking noise, occlusion and latency.
- **Higher demonstration throughput.** Across ten participants and five tasks, DexDirect collected more successful demonstrations, compared to TeleDex and AnyTeleop.
- **Lower operator workload.** In our user study, DexDirect reduced perceived mental demand, effort, and frustration, resulting in significantly lower overall workload than both baselines, as measured by NASA-TLX [17].

II. Related Work

A. Direct Physical Demonstration

One family of interfaces collects demonstrations by physically guiding a body that shares the robot's kinematics, so the trajectory is read directly from the joint encoders without pose estimation or retargeting. Kinematically isomorphic leaders use small-scale or duplicate robot arms as leader controllers, mapping joint space directly and thereby avoiding inverse kinematics failures near singularities while implicitly teaching operators the follower's physical limits [2, 3, 18, 19], and other works split the master device into large-range workspace navigation and fine-grained end-effector manipulation [20]. A key drawback, however, is that a second leader arm nearly doubles hardware requirements, resulting in high setup cost [21].

Other direct demonstration methods remove this duplicate setup by having the operator backdrive the deployment robot itself [22]. This also carries a cost: because the operator's hands occlude the camera's view of the contact region, each trajectory must be re-executed offline [23, 24], or the observation must be postprocessed through inpainting or point cloud prediction [25, 26].

B. Exoskeleton Interfaces

Exoskeleton interfaces instrument the operator's own limb and retarget the measured joint angles to the robot [7–9, 27–29], with several recovering contact force directly through the linkage. Related worn hand interfaces retarget the operator's finger motion while using visual inpainting [6] or a wrist camera [9] to obtain policy observations for the robot hand. These interfaces exhibit a common trade-off, with mechanical linkages, servo motors, or extra instrumentation that add further hardware, calibration, and per-operator fitting.

C. Vision and Pose-Based Teleoperation

Vision and pose-based teleoperation is a cheaper alternative to expensive master-device and exoskeleton systems [5, 10–13, 30, 31]. AnyTeleop [10] translates human movement to a robot using a standard webcam, pairing software-driven retargeting with a real-time motion generation module to map human keypoints to multi-fingered hands on the fly. TeleDex [12] further expands portability, turning a commodity smartphone into a standalone interface through a 3D-printable wrist mount for hand articulation, allowing for demonstration collection with minimal setup.

However, relying on monocular webcam vision leaves systems vulnerable to self-occlusion while inherently suffering from processing latency and tracking jitter [5]. While tracking a handheld device's own inertial and visual odometry circumvents visual occlusion [12], operating a phone in free space offers zero environmental resistance. This complete absence of haptic feedback leaves force regulation unconstrained, making continuous contact-rich manipulation, dynamic target tracking, and precise spatial coordination cognitively taxing and highly susceptible to operator error [32].

D. Handheld Capture

To enable real-time, single-pass collection, an alternative line of work decouples capture from the robot entirely using lower-cost wearable [32–34] or handheld interfaces [1, 35]. Several of these interfaces add synthetic haptic or wrench feedback so that operators can regulate contact force during collection [32, 34]. Although these devices can capture data in a single pass, their synthetic force feedback is governed by secondary control loops with inherent bandwidth limits, and recorded forces reflect proxy mechanics rather than the deployment manipulator's true joint inertia, motor compliance, and full wrench constraints [32, 35].

III. System Design

A. System Overview and Design Rationale

DexDirect separates arm motion and finger articulation into two complementary control pathways. As shown in

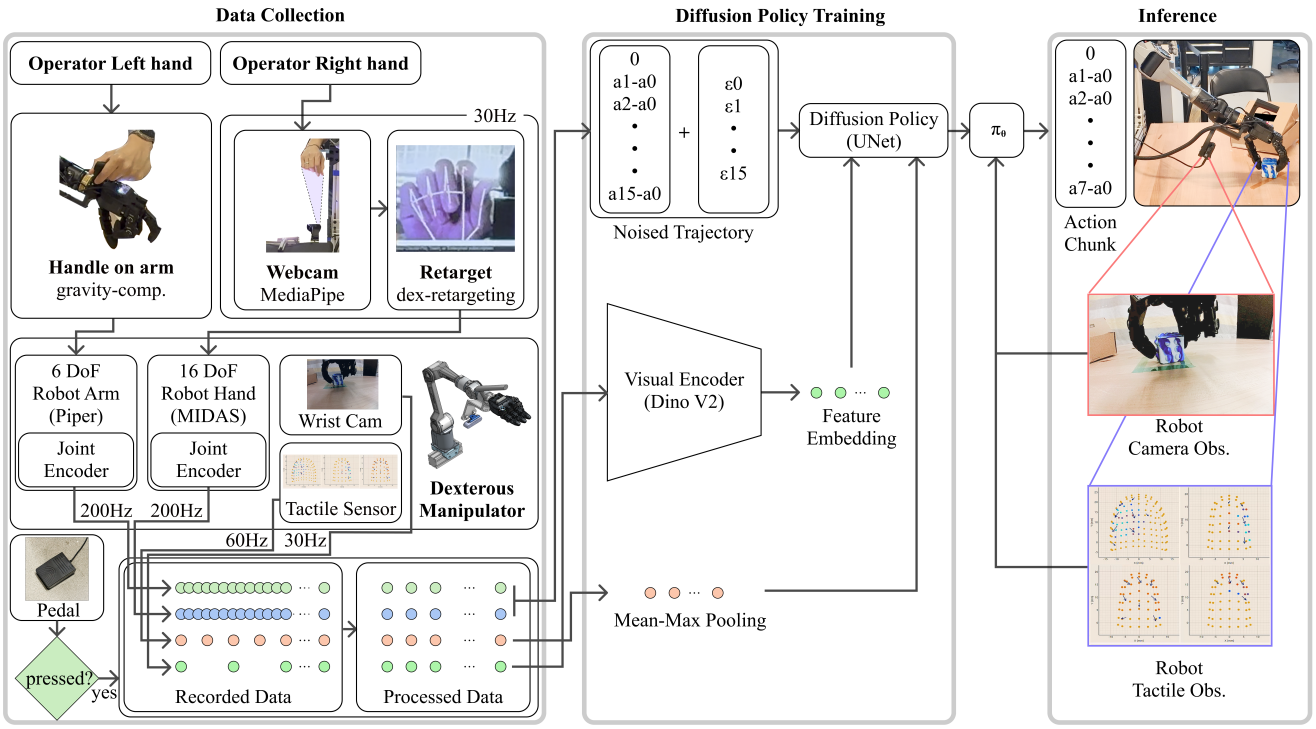


Fig. 2: **System overview of DexDirect** The operator directly guides the gravity-compensated robot arm while controlling the dexterous hand through vision-based hand-pose retargeting. The collected joint, visual, and tactile observations are synchronized and used to train a multimodal diffusion policy. At inference, the policy predicts robot action chunks conditioned on the current camera and tactile observations.

Fig. 2, the operator's left hand directly guides the physical robot arm through a rigidly attached handle, while the right hand controls the robot fingers through webcam-based hand tracking. A multimodal recording pipeline logs the resulting robot states, tactile measurements, and visual observations.

This separation follows from the different requirements of arm and finger control. Arm motion requires global 6-DoF positioning over a large workspace and involves physical interaction with the environment. Finger articulation is high-dimensional, but can be represented primarily by the relative configuration of the fingers in a hand-local frame, because the robot arm independently determines the global pose of the hand. DexDirect therefore uses direct kinesthetic guidance for the arm and vision-based retargeting for finger articulation. Because the arm determines the robot hand's global pose, the vision pathway uses only wrist-relative finger geometry, enabling a low-cost, non-wearable interface without arm-side pose tracking.

Manipulator. The platform consists of an Agilex PIPER 6-DoF manipulator and a dexterous MIDAS hand [36] mounted at the flange. The MIDAS hand has 16 joints driven by 13 actuators, is backdrivable, and contains 283 distributed tactile taxels. The total payload mounted at the end-effector flange is $0.75\,\mathrm{kg}$.

Operator handle. A simple 3D-printed PLA handle weighing $0.05\,\mathrm{kg}$ is mounted on the dorsal side of the hand assembly, with its grip point offset $65\,\mathrm{mm}$ from the flange along the palm normal. This offset provides clearance for the operator's hand while limiting the moment arm of the applied guidance force.

Sensing and operator station. A wrist-mounted RGB camera (Luxonis OAK-1 Wide FOV) records the visual observations used for policy learning. A separate webcam (Logitech C270), mounted $45\,\mathrm{cm}$ below the operator's right hand and oriented toward the palm, drives finger tracking. The operator stands beside the workspace with the left hand on the handle and the right forearm resting on a fixed 3D printed support. A nearby monitor displays the wrist-camera stream.

B. Direct Kinesthetic Arm Guidance

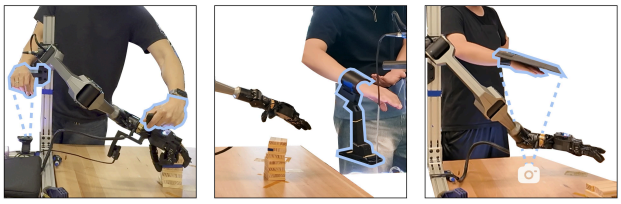
To compensate for the weight of the arm and end-effector payload while retaining low resistance to manual motion, the arm operates in a gravity-compensated drag mode with zero position stiffness and low joint damping. During demonstration collection, the joint-level controller runs at $200\,\mathrm{Hz}$. Before operation, we perform a staged system-identification procedure to estimate accurate gravity direction, end-effector payload mass and center of mass, joint-wise torque-delivery scaling, and Coulomb friction. These robot-specific parameters are used to correct the gravity torques computed from the MuJoCo model before applying them as feedforward torque commands. With zero position gain, the controller does not pull the arm toward a reference configuration, allowing the operator to directly position and orient the physical robot arm.

Since the physical robot itself is guided, arm trajectories are recorded directly from the joint encoders, and end-



Jenga extraction | USB unplugging | Lazy Susan: Pick & Place | Whiteboard Erasing | Keyboard Typing

Fig. 3: **Evaluation tasks.** DexDirect is evaluated on five contact-rich dexterous manipulation tasks: Jenga manipulation, USB unplugging, Lazy Susan pick-and-place, whiteboard erasing, and keyboard typing.



DexDirect | TeleDex | AnyTeleop

Fig. 4: **Comparison of operator interfaces for dexterous teleoperation.** DexDirect combines direct physical guidance of the robot arm with vision-based hand retargeting, whereas TeleDex relies on a handheld control interface and AnyTeleop tracks the operator's arm and hand motion entirely through vision.

effector poses are obtained through forward kinematics. No arm-side pose estimation, retargeting, inverse-kinematics solver, or Cartesian motion generator is required during collection. Each recorded configuration therefore corresponds to a state physically attained by the robot, subject to its joint and controller limits.

The rigid mechanical connection also transmits interaction force through the handle, providing the operator with passive contact cues without a separate haptic device. Joint motion remains bounded by the robot's hardware and controller safety limits, and an emergency stop is available throughout operation.

C. Vision-Based Finger Retargeting

A palm-facing webcam captures the operator's right hand at 30 Hz. Each RGB frame is processed by MediaPipe Hands [37], which detects 21 three-dimensional hand landmarks.

The wrist-relative landmarks are mapped to the MIDAS hand using the optimization-based `dex-retargeting` framework released with AnyTeleop [10]. At each time step, the optimizer computes robot joint targets by matching selected human and robot finger vectors subject to the robot joint limits.

No global wrist position or orientation estimated from the webcam is used for robot control, because the robot arm independently determines the global pose of the hand. This avoids relying on monocular metric wrist-pose estimation and confines the vision pipeline to hand-local finger articulation.

The forearm support helps keep the palm centered in the camera view and limits large wrist motion while leaving the fingers free to articulate. The tracked hand therefore does not need to reproduce the global motion or orientation of the robot hand.

D. Multimodal Demonstration Recording

Arm joint states, hand actuator states, tactile measurements, and wrist-camera images are timestamped using a common monotonic clock on the logging computer. Arm and hand states are logged at 200 Hz, tactile measurements at 60 Hz, and wrist-camera images at 30 Hz.

For policy training, the proprioceptive and tactile streams are aligned to the wrist-camera timestamps to form time-aligned 30 Hz demonstration episodes. For each camera timestamp, we select the nearest MIDAS hand encoder, PIPER end-effector pose, and tactile samples using nearest-neighbor timestamp matching, without temporal interpolation.

E. Policy Architecture

Our wrist-camera RGB frame is downscaled to 240×240, with a random 224×224 crop and color jitter during training and a center crop during inference. A DINOv2 ViT-S/14 encoder [38] turns this frame into an embedding that acts as the first part of the condition. In addition to the visual embedding, we append the 13-D vector of the absolute hand positions, along with a 4-D reading from the PaXini tactile sensors on the MIDAS Hand [36]: mean thumb, max thumb, mean index, and max index. The policy emits a 19-D action as its command: relative positions for the 6-DoF end-effector target and 13 finger targets.

IV. Experiments and Results

We evaluate DexDirect against two published low-setup teleoperation systems, AnyTeleop and TeleDex. Our evaluation addresses two primary questions: whether DexDirect increases the number of successful demonstrations collected under a fixed time budget, and whether it reduces the workload imposed on the operator. We additionally verify that demonstrations collected with DexDirect can be used for downstream policy learning.

A. Baseline Implementations

AnyTeleop. We reproduced the arm-control pathway of AnyTeleop on the same PIPER–MIDAS platform. An Intel RealSense RGB-D camera was used to estimate the operator's wrist pose following the pose-estimation procedure

TABLE I: Summary of quantitative results in the user study (mean ± SEM). Bold indicates the best performance for each metric.

	Jenga extraction		USB unplugging		Lazy Susan rotation and pick-and-place		Whiteboard erasing		Keyboard typing	
Method	**Success Rate**	**Time (s)†**	**Success Rate**	**Time (s)†**	**Success Rate**	**Time (s)†**	**Success Rate**	**Time (s)†**	**Success Rate**	**Time (s)†**
DexDirect	$\mathbf{0.94 \pm 0.02}$	$\mathbf{10.46 \pm 0.82}$	$\mathbf{0.95 \pm 0.02}$	$\mathbf{5.88 \pm 0.65}$	$\mathbf{0.96 \pm 0.02}$	$\mathbf{8.36 \pm 0.49}$	$\mathbf{0.91 \pm 0.04}$	$\mathbf{13.85 \pm 1.39}$	$\mathbf{0.71 \pm 0.04}$	$\mathbf{6.60 \pm 0.32}$
TeleDex	0.69 ± 0.08	49.76 ± 5.98	0.81 ± 0.04	10.36 ± 1.90	0.85 ± 0.06	21.96 ± 2.73	0.56 ± 0.12	39.10 ± 7.82	0.32 ± 0.06	17.50 ± 3.97
AnyTeleop	0.44 ± 0.12	35.50 ± 5.42	0.39 ± 0.09	17.75 ± 5.74	0.50 ± 0.13	47.14 ± 11.15	0.00 ± 0.00	—	0.00 ± 0.00	—

† Completion time was computed over successful task executions. A dash indicates that no successful executions were available for computing completion time.

described by AnyTeleop. The estimated wrist pose was transformed into a Cartesian end-effector target in the robot base frame. Hand-pose detection and end-effector target updates were performed at 30 Hz. Following AnyTeleop's CuRobo-based motion-generation pipeline, the target poses were passed to CuRobo's receding-horizon `MpcSolver`, which ran at 120 Hz on a separate workstation equipped with an NVIDIA RTX 4070 Ti. The resulting joint-space trajectories were streamed to the robot's low-level controller. To improve hand tracking reliability, the operator wore a black lightweight plate on the back of the hand.

TeleDex. We implemented the TeleDex baseline using the authors' publicly released open-source software and hardware.

B. User Study Design

Participants and interfaces. We conducted a user study with ten participants who had no prior experience with DexDirect and both baseline methods. Each participant operated the same PIPER–MIDAS robot using three interfaces: DexDirect, TeleDex, and AnyTeleop. Across conditions, the robot arm, dexterous hand, finger-retargeting pipeline, visual feedback, and task environments were held fixed; the interfaces differed in how the operator commanded the global motion of the robot hand.

The interface order was counterbalanced across participants using the cyclic orders ABC, BCA, and CAB. Each participant completed the full five-task sequence with one interface before proceeding to the next interface.

Before the timed evaluation of each task, participants were given one minute to practice the task with the corresponding interface. For each task, participants were instructed to complete as many successful repetitions as possible within a fixed time budget. The budget was 120 s for each task except Jenga, for which participants received 180 s. Each completed attempt was classified as successful or unsuccessful according to the task-specific criteria described below.

Tasks. The five tasks shown in Fig. 3 were selected to span distinct arm-control requirements, including large-workspace positioning, sustained contact-force regulation, discrete precision contact, and constrained extraction.

- *Lazy Susan rotation and pick-and-place* evaluates sustained tangential contact, workspace reconfiguration, and global arm positioning. The target cube is initially placed on the side of the platform farthest from the operator and robot hand. The participant first rotates the Lazy Susan by approximately 180° through direct contact to bring the cube into the near workspace, then grasps the cube and places it in the designated target region. A repetition is successful when the platform is rotated as prescribed and the cube is transferred to the target region without being dropped.
- *Whiteboard erasing* requires simultaneous tangential motion along the board and regulation of normal contact force. The participant erases the designated marking while maintaining contact between the eraser and the board. A repetition is successful when the word 'robot' in the prescribed region is fully erased.
- *Keyboard typing* evaluates repeated discrete contact with small spatial targets. The participant presses the prescribed key sequence 'bot'. A repetition is successful when all requested keys are registered in the correct order without an incorrect key press.
- *Jenga extraction* combines precise hand placement, stable fingertip contact, and gradual regulation of extraction force. The participant pinches and completely removes the designated block. A repetition is successful when the target block is extracted without collapsing the stack.
- *USB unplugging* requires a stable grasp, orientation alignment with the connector axis, and directed extraction force. A repetition is successful when the connector is fully removed from the socket.

An attempt was marked as unsuccessful if the success criterion was not met, an irreversible task error occurred, or the allotted time elapsed before completion.

C. Demonstration Throughput and Task Performance

Under the same time budget per-condition, DexDirect collected 481 successful demonstrations across participants and tasks, compared with 151 for TeleDex and 28 for AnyTeleop. This corresponds to $3.2\times$ the successful-demonstration throughput of TeleDex and $17.2\times$ that of AnyTeleop.

As summarized in Table I, DexDirect achieved the highest success rate on every task. Its success rate ranged from 0.71 on keyboard typing to 0.96 on Lazy Susan rotation and pick-and-place, whereas TeleDex ranged from 0.32 to 0.85 and AnyTeleop from 0 to 0.50. In particular, AnyTeleop produced no successful whiteboard or keyboard demonstrations within the evaluation budget. Video review showed that participants struggled to regulate the end-effector depth normal to the whiteboard, resulting in insufficient or unstable contact during erasing. During keyboard typing, the delayed robot response made brief press-and-release motions difficult to

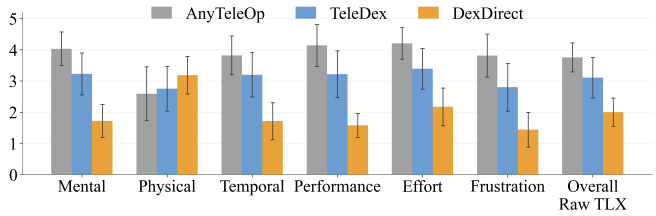


Fig. 5: Mean adapted NASA–TLX scores for AnyTeleop, TeleDex, and DexDirect, averaged across ten participants and five tasks. Error bars indicate 95% confidence intervals computed from participant-level scores. Lower scores indicate lower perceived workload. The performance item is reverse-coded so that higher values consistently indicate poorer perceived performance.

execute, frequently causing repeated registrations of the same key.

DexDirect also reduced the time required for successful task executions. Relative to TeleDex, it achieved completion-time speedups of $4.8\times$ on Jenga, $1.8\times$ on USB unplugging, $2.6\times$ on Lazy Susan, $2.8\times$ on whiteboard erasing, and $2.7\times$ on keyboard typing. Relative to AnyTeleop, the speedups on tasks for which AnyTeleop produced successful trials were $3.4\times$ on Jenga, $3.0\times$ on USB unplugging, and $5.6\times$ on Lazy Susan.

These results indicate that DexDirect's throughput advantage is not caused solely by allowing more unsuccessful attempts. Participants both completed individual trials more quickly and succeeded on a larger fraction of those attempts.

D. Operator Workload

DexDirect substantially reduced subjective workload despite requiring the operator to physically move the robot arm. The adapted Raw TLX score is the unweighted mean of six workload dimensions, with lower scores indicating lower perceived workload. The performance item was reverse-coded so that higher values consistently indicate poorer perceived performance and greater workload. DexDirect achieved an overall Raw TLX score of 2.01, compared with 3.08 for TeleDex and 3.81 for AnyTeleop. Paired Wilcoxon signed-rank tests on participant-level scores found a significant reduction relative to both AnyTeleop ($p = 0.00195$) and TeleDex ($p = 0.01953$). Both values are below the conventional significance threshold of 0.05, indicating that the observed reductions are unlikely to arise from random participant-level variation alone.

The largest differences occurred in cognitive dimensions. Relative to TeleDex, DexDirect reduced mental demand from 3.19 to 1.77, frustration from 2.79 to 1.47, and effort from 3.40 to 2.17. Temporal demand and perceived-performance workload were also lower. Physical demand was the only dimension for which DexDirect performed worse, increasing from 2.72 for TeleDex and 2.62 for AnyTeleop to 3.23.

The workload profile therefore exposes the central trade-off of direct kinesthetic guidance: it replaces cognitive effort associated with tracking, coordinate remapping, and visual correction with the physical effort of moving the gravity-compensated robot. The lower overall workload suggests that this trade is favorable for repeated demonstration collection.

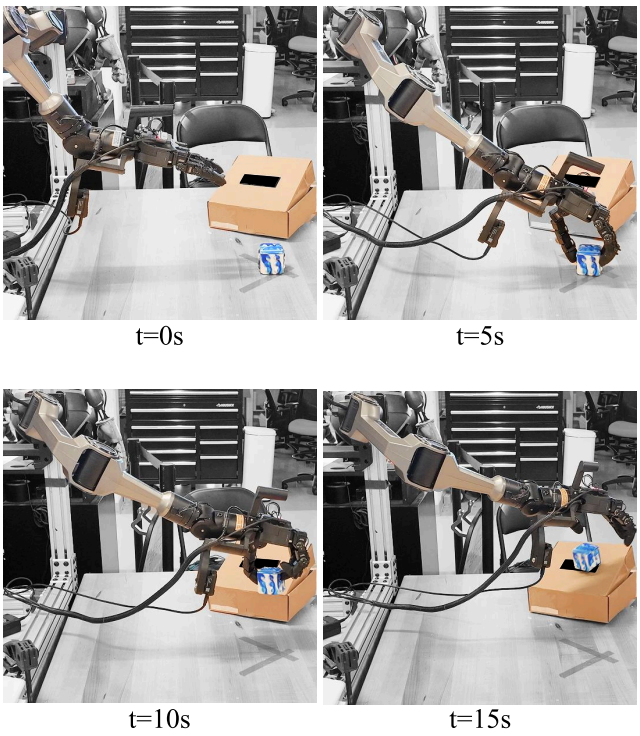


Fig. 6: **Representative policy rollouts.** The learned policy grasps a cube from a randomized initial pose, lifts it, and places it onto the target box. Each sequence shows the rollout at $t = \{0, 5, 10, 15\}$ s.

E. Policy Training

To evaluate whether DexDirect is suitable for end-to-end policy learning, we design a simple cube pick-and-place task. The robot picks up a cube from the table and drops it onto a box located $30\,\mathrm{cm}$ away. Success is defined as lifting the cube, transporting it above the box, and releasing it such that it lands on the box.

For policy learning, a single operator separately collects 200 demonstrations using DexDirect. These demonstrations are collected independently of the user study. We train a diffusion policy for the cube pick-and-place task and evaluate the policy over 20 episodes without operator intervention, using the success definitions described above.

The cube pick-and-place task achieves a success rate of 90% (18/20) over 20 evaluation episodes.

V. Limitations

DexDirect improves demonstration throughput through direct physical arm guidance, but introduces several trade-offs.

First, the system is inherently co-located and requires a robot capable of safe, low-impedance, gravity-compensated guidance. It therefore does not support remote teleoperation and may be less ergonomic for larger or higher-inertia manipulators. Consistent with this trade-off, DexDirect increased physical demand despite reducing overall workload.

Second, the current interface uses both operator hands to control one arm-hand system, preventing direct single-operator bimanual control. A possible extension is to mechanically couple each operator hand to one robot arm while tracking the fingers of the same hand, although its visibility, ergonomics, and safety remain to be studied.

Third, finger control relies on monocular landmark detection and morphology-dependent retargeting, and is therefore still susceptible to tracking and retargeting errors. The interface also does not provide distributed fingertip force feedback.

Finally, our evaluation is limited to ten participants, five short tasks, one robot platform, and a policy-learning task. It also evaluates the combined effect of direct coupling without isolating coordinate remapping, latency, and passive contact cues. Future work will consider longer-term studies, additional platforms, bimanual operation, and controlled ablations of these factors.

VI. Conclusion

We presented DexDirect, a low-setup interface that combines direct kinesthetic arm guidance with vision-based finger retargeting for efficient dexterous demonstration collection. In a ten-participant study across five tasks, DexDirect collected 481 successful demonstrations, compared with 151 using TeleDex and 28 using AnyTeleop, while reducing completion time and overall workload. Policy trained on DexDirect demonstrations achieved 90% success on cube pick-and-place. These results show that eliminating the arm-side tracking and retargeting loop provides a practical and effective path toward scalable dexterous robot learning.

References

[1] C. Chi et al., "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Robotics: Science and Systems (RSS)*, 2024. arXiv: 2402.10329 [cs.RO].

[2] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024. arXiv: 2309.13037 [cs.RO].

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023. arXiv: 2304.13705 [cs.RO].

[4] Z. Si et al., *Exostart: Efficient learning for dexterous manipulation with sensorized exoskeleton demonstrations*, 2025. arXiv: 2506.11775 [cs.RO].

[5] A. Handa et al., "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[6] M. Xu et al., "Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation," in *Conference on Robot Learning*, PMLR, 2025, pp. 437–459.

[7] S. Yang et al., "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," in *Conference on Robot Learning (CoRL)*, 2024. arXiv: 2408.11805 [cs.RO].

[8] H.-S. Fang et al., *Dexop: A device for robotic transfer of dexterous human manipulation*, 2025. arXiv: 2509.04441 [cs.RO].

[9] A. Zhu et al., *Dexexo: A wearability-first dexterous exoskeleton for operator-agnostic demonstration and learning*, 2026. arXiv: 2603.17323 [cs.RO].

[10] Y. Qin et al., "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," in *Robotics: Science and Systems*, 2023.

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

[12] O. Rayyan, M. Gilles, and Y. Cui, *Teledex: Accessible dexterous teleoperation*, 2026. arXiv: 2603.17065 [cs.RO].

[13] T. He et al., "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Conference on Robot Learning (CoRL)*, 2024. arXiv: 2406.08858 [cs.RO].

[14] S. Wrede, C. Emmerich, R. Grünberg, A. Nordmann, A. Swadzba, and J. J. Steil, "A user study on kinesthetic teaching of redundant robots in task and configuration space," *Journal of Human-Robot Interaction*, vol. 2, no. 1, pp. 56–81, 2013.

[15] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz, "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, ACM, 2012, pp. 391–398.

[16] H. Li, Y. Cui, and D. Sadigh, "How to train your robots? the impact of demonstration modality on imitation learning," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 1113–1120.

[17] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., North-Holland, 1988, pp. 139–183.

[18] R. Cadene et al., *Lerobot: State-of-the-art machine learning for real-world robotics in pytorch*, https://github.com/huggingface/lerobot, 2024.

[19] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conference on Robot Learning (CoRL)*, 2024. arXiv: 2401.02117 [cs.RO].

[20] J. Zhou, B. Liang, J. Huang, I. Zhang, and M. Tomizuka, *Global-local interface for on-demand teleoperation*, 2025. arXiv: 2502.09960 [cs.RO].

[21] A. Iyer et al., *Open teach: A versatile teleoperation system for robotic manipulation*, 2024. arXiv: 2403.07870 [cs.RO].

[22] 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.

[23] C. Chen, Z. Yu, H. Choi, M. Cutkosky, and J. Bohg, "Dexforce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation," *IEEE Robotics and Automation Letters (RA-L)*, 2025. arXiv: 2501.10356 [cs.RO].

[24] T. Ablett et al., "Multimodal and force-matched imitation learning with a see-through visuotactile sensor," *IEEE Transactions on Robotics (T-RO)*, vol. 41, pp. 946–959, 2025. arXiv: 2311.01248 [cs.RO].

[25] D. Zhang, C. Yuan, C. Wen, H. Zhang, J. Zhao, and Y. Gao, *Kinedex: Learning tactile-informed visuomotor policies via kinesthetic teaching for dexterous manipulation*, 2025. arXiv: 2505.01974 [cs.RO].

[26] U. Yoo, J. Francis, J. Oh, and J. Ichnowski, "Kinesoft: Learning proprioceptive manipulation policies with soft robot hands," in *Conference on Robot Learning (CoRL)*, 2025. arXiv: 2503.01078 [cs.RO].

[27] H. Fang et al., "Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons," in *Conference on Robot Learning (CoRL)*, 2025. arXiv: 2503.03081 [cs.RO].

[28] L. Liang et al., *Universal manipulation exoskeleton: Learning compliant whole-body policies with real-time torque feedback*, 2026. arXiv: 2606.14218 [cs.RO].

[29] R. Zhong et al., *Nuexo: A wearable exoskeleton covering all upper limb rom for outdoor data collection and teleoperation of humanoid robots*, 2025. arXiv: 2503.10554 [cs.RO].

[30] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," in *Conference on Robot Learning (CoRL)*, 2024. arXiv: 2407.01512 [cs.RO].

[31] R. Ding et al., *Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning*, 2024. arXiv: 2407.03162 [cs.RO].

[32] Y. Gao, H. Ma, and P. Zheng, *Glovity: Learning dexterous contact-rich manipulation via spatial wrench feedback teleoperation system*, 2025. arXiv: 2510.09229 [cs.RO].

[33] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," in *Robotics: Science and Systems (RSS)*, 2024. arXiv: 2403.07788 [cs.RO].

[34] J. Koh, H. Jung, N. Kim, W. Ko, and C. Nam, *Dex-mouse: A low-cost portable and universal interface with force feedback for data collection of dexterous robotic hands*, 2026. arXiv: 2604.15013 [cs.RO].

[35] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu, "Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. arXiv: 2410.07554 [cs.RO].

[36] A. Zhu, M. Zhu, B. J. Kim, Q. Wang, J. V. S. H. Ramos, and D. Hong, *Midas hand: Modular low-impedance direct-drive anthropomorphic sensing hand*, 2026. arXiv: 2607.14487 [cs.RO].

[37] F. Zhang et al., *Mediapipe hands: On-device real-time hand tracking*, 2020. arXiv: 2006.10214 [cs.CV].

[38] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv: 2304.07193 [cs.CV].