SEED-UMI: Sharing the Exoskeleton between human and robot for onE-to-one Dexterous demonstration

Tengbo Yu [1,2,*] **Jiahao Wu** [2,*] **Daohan Li** [2,*] **Bingxu Chen** [2]
Hao Liu [2] **Xiaojian Ma** [2] **Hangxin Liu** [1,†]
[1] State Key Laboratory of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[2] Delta Intelligence
https://tengbo-yu.github.io/SEED-UMI/

Abstract: Imitation learning for dexterous hands is bottlenecked by the difficulty of collecting contact-rich demonstrations that transfer faithfully to the robot. Prior wearable-exoskeleton systems record only on the human side and retarget via open-loop mappings calibrated in free space, which degrade under contact. We present **SEED-UMI**, a framework in which *both the human and the robot wear the same exoskeleton*: joint encoders become a physically shared measurement, and wrist cameras mounted to the exoskeleton observe the same outer mechanism during both human data collection and robot policy rollouts. This turns retargeting into paired cross-embodiment supervision and lets policies train directly on raw exoskeleton-centric wrist images, without segmentation or inpainting. On five contact-rich tasks, SEED-UMI achieves $3.0\times$ greater data collection efficiency than exoskeleton-based teleoperation and a 70.0% average rollout success rate.

Keywords: Dexterous Manipulation, Wearable Exoskeleton, Demonstration Data Collection, Cross-Embodiment Supervised Learning

1 Introduction

Dexterous hand manipulation remains one of the most challenging problems in robot learning due to the high degrees of freedom of robotic hands and the complex contact-rich interactions involved during manipulation. Although imitation learning (IL) has enabled robots to acquire sophisticated locomotion and arm manipulation skills, progress in dexterous manipulation has been comparatively limited. A major bottleneck lies in demonstration collection and retargeting [1, 2, 3, 4]: accurately capturing natural human hand motions while transferring them reliably to robot remains difficult because human and robot embodiments differ in kinematics, sensing, and contact behavior [5].

Existing demonstration collection approaches each involve a trade-off between accurately recording dexterous hand motions and preserving natural hand behavior, which are often competing. Vision-based hand tracking allows demonstrations to be collected without constraining hand motion [6, 7, 8, 9], but data quality would degrade due to occlusion and temporal jitter when contacting with external objects. Glove-based systems equipped with inertial or flex sensors [10, 11, 12, 13] provide more reliable motion data for interactive tasks, but are prone to skin-motion artifacts and require complex correspondence mapping between human and robot joints. Exoskeleton-based devices produce the most consistent hand gesture measurements that are identical to mechanically constrained hand motions in the robot side, yet often sacrifice operator comfort and demonstration throughput due to bulky and restrictive hardware [14, 15].

More importantly, these collection approaches fundamentally shape the downstream learning problem. Vision-based systems typically simplify dexterous manipulation into end-effector control be-

[*] Indicates equal contribution.
[†] Corresponding author. Email: hx.liu@pku.edu.cn

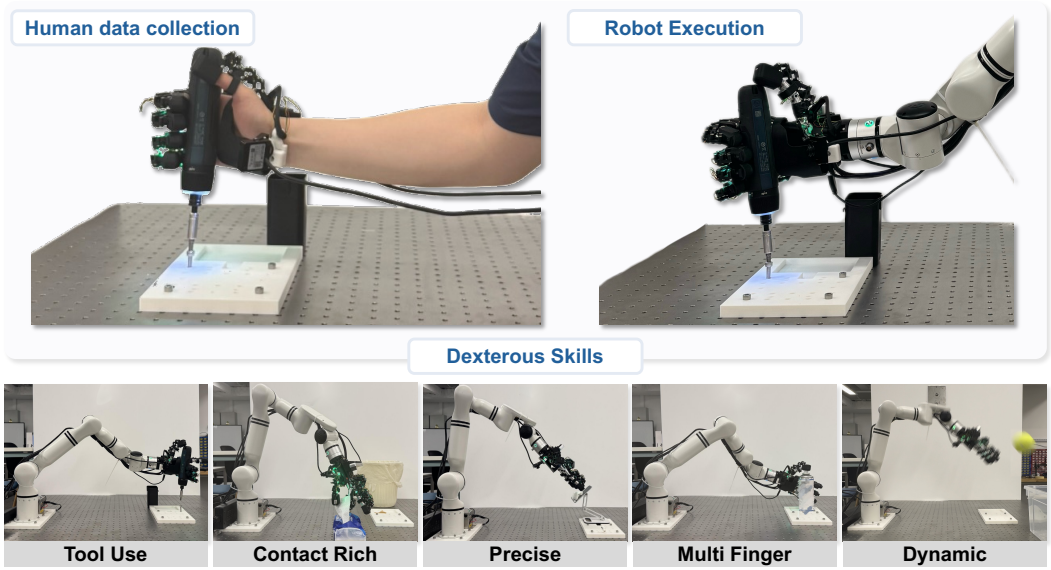

Figure 1: **SEED-UMI data-collection interface and task suite.** *Top:* the human operator and the dexterous robot hand wear the same exoskeleton, so demonstrations and robot replay are recorded through a shared physical device with matched encoders and wrist-mounted cameras. *Bottom:* representative contact-rich tasks enabled by the interface, spanning tool use, small-object placement, grasp-and-throw timing, contact-rich table cleaning, and multi-finger spray actuation.

cause fine-grained finger motions are more difficult to recover reliably [16]. Glove-based methods require extensive post-processing and retargeting to establish correspondence with robot hand kinematics [17, 18]. As a result, current dexterous imitation learning pipelines largely treat human demonstrations and robot execution as two separate domains connected through increasingly complex retargeting models.

Recently, exoskeleton-based systems have attracted renewed interest because they offer a promising way to reduce this correspondence gap through co-design with the target robot hand [19, 20, 21]. By aligning joint encoders with robot degrees of freedom and mounting wrist cameras from the robot end-effector viewpoints, such systems enable learning of more dynamic and reactive manipulation behaviors. However, existing approaches still operate as one-way mappings from human motion to robot motion. They typically calibrate retargeting in free space [19], which often degrades under contact-rich interactions due to friction, external loads, and joint coupling effects. Furthermore, systems such as DexUMI [19, 20, 21] rely on generative image translation to bridge visual discrepancies between human and robot observations, while robot-side measurements themselves are not used to improve correspondence learning.

In this paper, we present **SEED-UMI**, a demonstration-collection and policy-learning framework built upon the idea of *shared physical measurement*. Instead of learning a mapping between separate human and robot hand embodiments, we design a shared measurement interface that can be directly used by both. Specifically, we develop a kinematically isomorphic exoskeleton co-designed with the target robot hand, whose link lengths, joint axes, and motion ranges are jointly optimized such that the same exoskeleton device can be worn by both the human hand and the robot hand. As a result, both human demonstrations and robot executions generate encoder measurements under the same mechanical structure across all 20 degrees of freedom (Section 3).

Building upon this shared interface, we learn a contact-aware encoder-to-command mapping through a two-stage process consisting of motor babbling followed by paired trajectory fine-tuning (Section 4). Importantly, discrepancies between human and robot encoder trajectories under the same

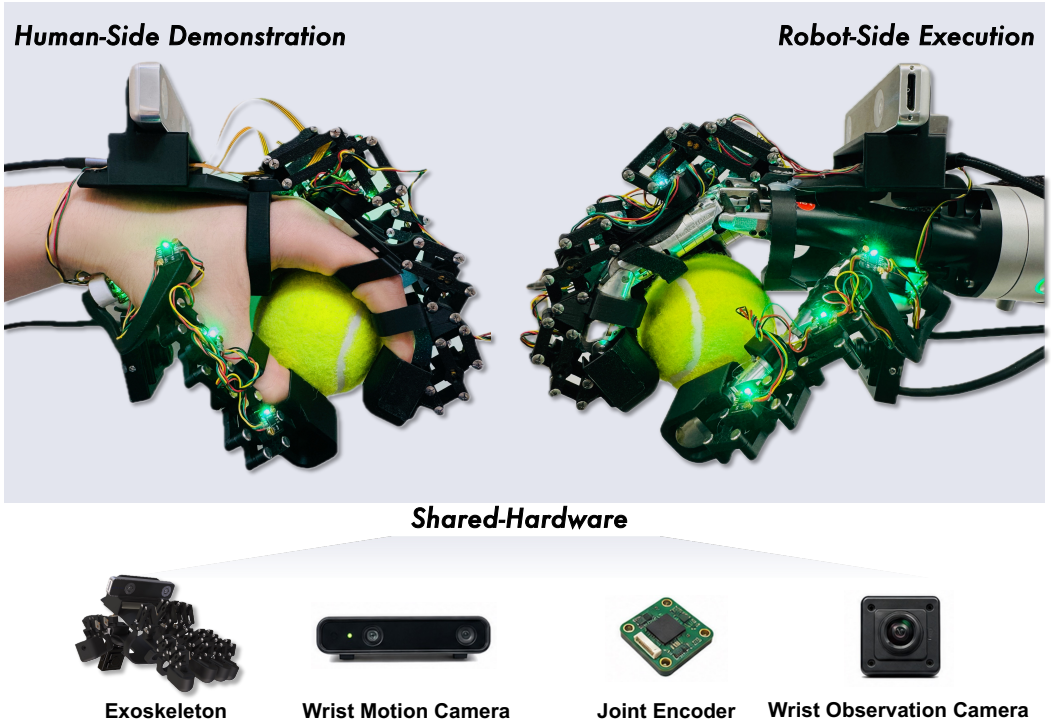

Figure 2: **Shared exoskeleton hardware.** The same skeleton is worn by the human operator (*left*) and the fully-actuated target hand (*right*). One contactless rotary encoder per joint, a dorsal T265 for 6-DoF wrist tracking, and a ventral fisheye camera are rigidly mounted on the exoskeleton frame, so the encoder vector $\mathbf{e}_t$ and the wrist-camera intrinsics are identical across embodiments.

intended motion naturally become paired supervisory signals for retargeting and policy learning. In addition, policies are trained directly from raw exoskeleton-centric wrist observations, avoiding post-hoc segmentation and generative image translation pipelines.

Combining these components, SEED-UMI enables natural, contact-rich demonstrations collected from human operators to be transferred directly to robot execution, decouples human demonstration collection from real-time robot teleoperation, simulation, or reinforcement learning. We evaluate our framework on five dexterous manipulation tasks spanning tool use, precise object placement, grasp-and-throw timing, contact-rich table cleaning, and multi-finger spray actuation (Section 5). SEED-UMI achieves a data collection speed 3.0$\times$ faster than exoskeleton-based teleoperation. With paired fine-tuning, learned policies achieve a mean success rate of 70.0% across all tasks, approaching the 71.7% teleoperation baseline while outperforming it on highly precise and dynamic tasks such as throwing.

Our results suggest that overcoming the demonstration collection and retargeting bottleneck in dexterous manipulation may require rethinking the interface between human and robot embodiments rather than solely improving retargeting methods. By *transforming retargeting and visuomotor policy learning into a form of paired cross-embodiment supervision*, our idea of shared physical measurement offers a promising direction toward more scalable and efficient dexterous robot learning.

2 Related Work

Wearable exoskeletons for dexterous demonstration. Glove-based systems [13, 11, 12, 10, 18, 15, 22] record human hand pose but lack one-to-one joint correspondence with the target robot. Wearable [19, 23, 24, 25, 26] constrain the hand into a robot-matched kinematic structure: DexUMI [19] instruments the exoskeleton with joint encoders and adapts it to multiple robot hands;

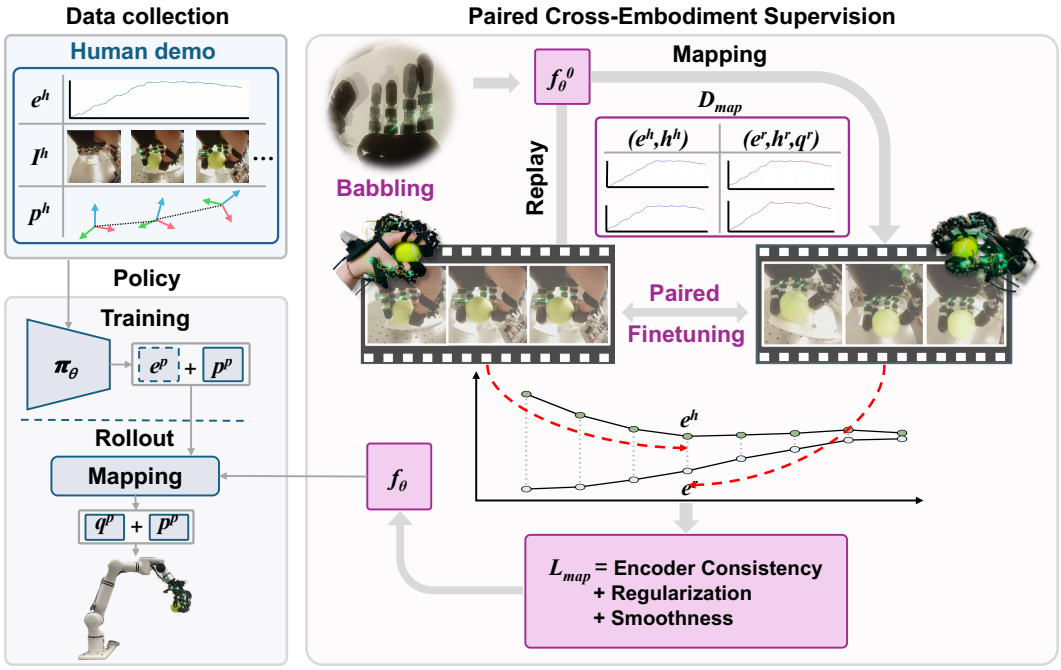

Figure 3: **SEED-UMI pipeline.** Human demonstrations are collected through the shared exoskeleton, paired robot replay refines the encoder-to-command mapping, and the learned policy rolls out on the robot with matched exoskeleton-centric observations.

WHED [23] improves wearability with a pose-tolerant thumb mechanism. These systems mount the exoskeleton on the human side only and transfer demonstrations via a one-shot free-space regression. SEED-UMI extends this paradigm by mounting the same exoskeleton on the robot, converting the open-loop mapping into a supervised learning problem with paired cross-embodiment data.

Visual alignment for cross-embodiment policy learning. Inpainting-based methods [20, 21] and the visual pipeline of DexUMI [19] bridge the appearance gap between embodiments by hallucinating the target hand in the observation stream. These operate at the policy interface and require generative models. We instead remove most of the appearance gap by design: because both embodiments wear the same exoskeleton and the wrist camera is rigidly mounted to that exoskeleton, data collection and policy rollouts expose the policy to the same outer mechanism, object, and contact interface. SEED-UMI therefore trains on raw wrist images with standard visual augmentation, without hand segmentation or generation.

Encoder-to-motor mapping. DexUMI [19] fits a per-joint regression in free space, which degrades under contact load on fully-actuated hands. DexterityGen [27, 3, 28, 29, 30, 31] trains a low-level controller via large-scale simulation RL. Our approach occupies a middle ground: we obtain real-world supervision by replaying demonstrations on the robot through the shared exoskeleton, requiring neither simulation nor reinforcement learning, in the spirit of demonstration-augmentation methods [32, 33, 34, 35] but specialized to the encoder-to-command mapping.

3 Hardware: A Shared Exoskeleton

The hardware of SEED-UMI matches the contact-facing exoskeleton geometry and encoder frame on the human and robot sides, while adapting cuffs and mounts to each embodiment. This makes encoder readings a shared *cross-embodiment measurement*: discrepancies under the same command supervise the residual command-to-joint mapping (§4). Figure 2 shows both systems.

3.1 Exoskeleton Mechanism Design

The exoskeleton is co-designed with a single fully-actuated dexterous hand as its sole target platform. Because we do not pursue cross-platform transferability, we can fix link lengths, joint axes, and range-of-motion limits to a direct mechanical replica of the robot's kinematic tree, so that we can enforce two stronger properties: 1:1 joint correspondence on *both* sides of the device, and a contact geometry that is identical between the human-worn and robot-worn instances.

E.1 Isomorphic kinematic skeleton. The skeleton has one revolute joint per active degree of freedom of the target hand, with link lengths matched to the robot's phalanx lengths and joint axes oriented identically. Hard stops mirror the robot's range-of-motion limits, so configurations the operator can reach are exactly those the robot can reach. The exoskeleton instantiates 20 independently measured joints across five fingers.

E.2 Parallel four-bar linkage for encoder routing. A naive layout that places the encoder directly on the lateral interphalangeal axis is mechanically simple but fails two requirements at once: the housing protrudes into the lateral clearance envelope and restricts abduction/adduction below the robot's native range; it also occludes part of the palmar contact surface, breaking the shared-geometry property between embodiments. We therefore route the rotation through a *parallel four-bar linkage* to a dorsally mounted contactless magnetic encoder (Figure 4 (b)). The linkage preserves one-to-one angular correspondence between the joint axis and encoder shaft, while the palmar and lateral surfaces remain unobstructed (Figure 4 (a)). Link lengths are tuned per finger to minimize dorsal height, and press-fit magnet pockets with precision sleeve bearings at each pivot keep encoder alignment stable across repeated wear cycles.

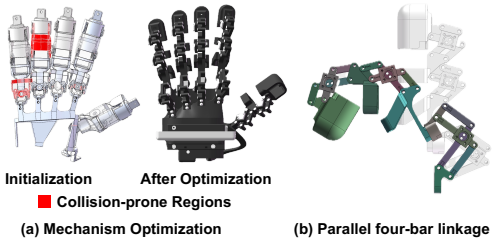

Figure 4: **Hardware Design.** (a) shows the collision avoidance of our design during operation. (b) shows how the linkage work during operation.

E.3 Identical contact geometry on both sides. The human- and robot-worn instances share the same outer skeleton: link lengths, joint axes, dorsal linkage, and palmar fingertip pads. Cuffs and mounts adapt this skeleton to each embodiment without changing joint frames. Thus both embodiments present the same external contact geometry to the object, and human–robot encoder discrepancies mainly reflect load-dependent effects such as contact force, friction, and hysteresis, which provide the residual signal used in §4.

3.2 Sensor Integration

All sensing is mounted on the exoskeleton frame so that every signal is recorded in the same physical frame regardless of which embodiment wears the device. Two design objectives drive the sensor layout: (i) capture the action and observation channels needed for downstream policy learning, and (ii) minimize the distribution shift of those channels between the human-side and robot-side.

S.1 Joint encoders. To precisely capture joint actions, our exoskeleton integrates contactless magnetic rotary encoders at every actuated joint on *both* embodiments—the human-worn and the robot-worn instances share the same encoder placement. Concatenating per-joint readings yields $\mathbf{e}_t \in \mathbb{R}^{20}$. Due to the four-bar linkage routing and manufacturing tolerances, the mapping between exoskeleton encoder readings $\mathbf{e}_t$ and robot joint commands $\mathbf{q}_t$ is non-linear; we therefore learn this mapping with a data-driven model described in §4.

S.2 Wrist-mounted cameras. A dorsal Intel RealSense T265 records 6-DoF wrist pose, while a ventral fisheye camera (150° FoV) observes the workspace. Both are rigidly mounted to the exoskeleton, fixing their intrinsics and encoder-frame extrinsics across the human and robot sides.

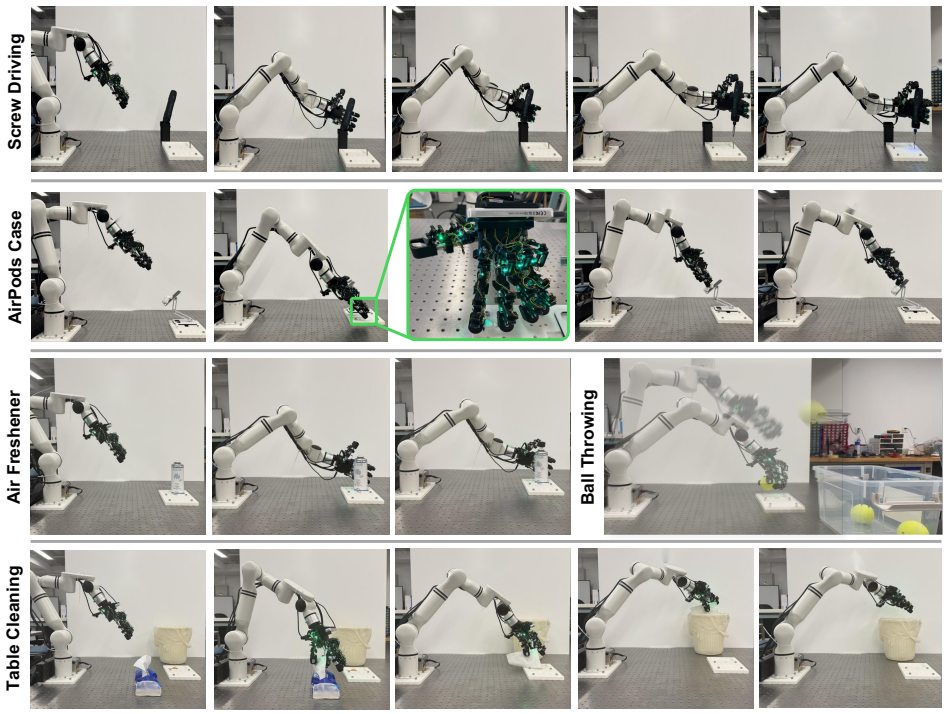

Figure 5: **Policy Rollouts.** We evaluate SEED-UMI on five contact-rich dexterous manipulation tasks that together stress tool stabilization, index–middle finger coordination, grasp-to-release timing, contact-rich table wiping, and thumb-involved multi-finger actuation.

Because data collection and rollout use the same exoskeleton-camera assembly, the policy receives aligned raw observations I_t with only standard photometric and crop augmentations. Unlike DexUMI [19], this hardware alignment requires no hand segmentation or inpainting, preserving contact pixels for fine manipulation.

4 Software Adaptation: Two-Stage Cross-Embodiment Mapping

The shared exoskeleton gives both embodiments a common encoder signal $\mathbf{e}_t \in \mathbb{R}^{20}$, but robot execution still requires an encoder-to-command mapping $\mathbf{e}_t \mapsto \mathbf{q}_t$. Instead of fitting this mapping once in free space [19], we bootstrap it with robot motor babbling and refine it with paired replay under real contact. The mapping is proprioceptive, while aligned wrist images are reserved for policy learning; Figure 3 shows the full pipeline.

M.1 Motor babbling on the robot. The robot wears the exoskeleton and executes a structured exploration protocol, covering single-joint sweeps, coupled motions at varied speeds, and a few self-contact poses. We record robot joint positions $\mathbf{q}_t$ and the robot-side encoder reading $\mathbf{e}_t^{(r)}$, the history vector $\mathbf{h}_t$ concatenates the previous K encoder readings, finite-difference encoder velocities, and a binary opening/closing direction indicator for each joint, and fit an initial mapping $f_{\theta^0}\colon (\mathbf{e}_t, \mathbf{h}_t) \mapsto \mathbf{q}_t$. This free-space model captures basic kinematics and history effects, but still drifts under contact, motivating the paired replay stage in M.2.

M.2 Paired replay and fine-tuning. The human operator performs representative object-grasping motions while wearing the exoskeletons, recording contact-rich encoder trajectories $\{\mathbf{e}_t^{(h)}, \mathbf{h}_t^{(h)}\}$. We replay each demonstration on the robot through f_{θ^0} and record $\{\mathbf{e}_t^{(r)}, \mathbf{h}_t^{(r)}, \mathbf{q}_t^{(r)}\}$ from robot-side exoskeleton. This produces a paired mapping dataset $\mathcal{D}_{\text{map}} = \{(\mathbf{e}_t^{(h)}, \mathbf{h}_t^{(h)}, \mathbf{q}_t^{(r)}, \mathbf{e}_t^{(r)}, \mathbf{h}_t^{(r)})\}$, with a forward model g_ϕ from robot commands and encoder history to robot-side encoder readings,

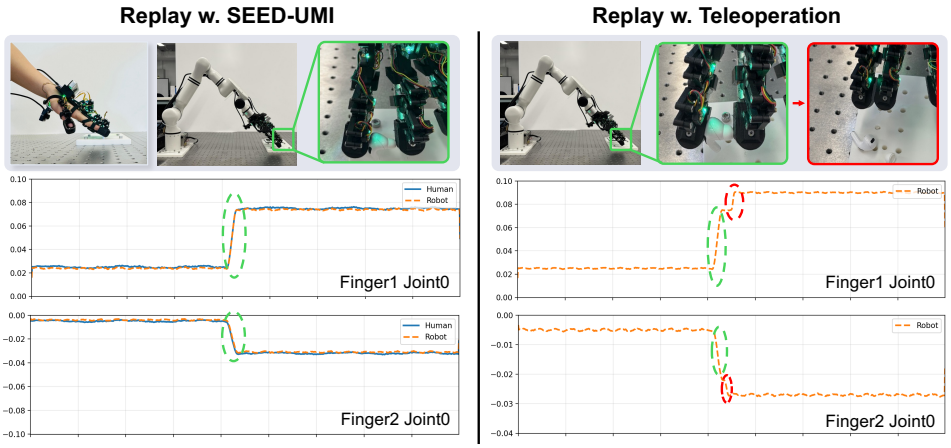

Figure 6: **Fine manipulation comparison.** AirPods task with human- (blue) and robot-side (orange) encoder traces.

and fine-tune f_θ as an inverse problem:

$$\mathcal{L}_{\text{map}} = \|g_\phi(f_\theta(\mathbf{e}_t^{(h)}, \mathbf{h}_t^{(h)}), \mathbf{h}_t^{(h)}) - \mathbf{e}_t^{(h)}\|_2^2 + \beta\|f_\theta(\mathbf{e}_t^{(h)}, \mathbf{h}_t^{(h)}) - \mathbf{q}_t^{(r)}\|_2^2 + \lambda\|\Delta f_\theta\|_2^2, \quad (1)$$

The loss matches predicted robot-side encoders to the human trace, regularizes commands toward paired replay, and penalizes temporal jumps. We can repeat replay-and-refine once to turn the free-space babbling model into a contact-robust encoder-to-command function without manual retuning.

The paired replay stage does not assume that the initial replay perfectly reproduces the human trajectory. Instead, it uses the discrepancy between the human-side encoder trace and the robot-side encoder response as a correction signal for the command mapping.

M.3 Policy training and rollout. For downstream policy training we adopt the ACT [36], Diffusion Policy (DP) [37] and $\pi_{0.5}$ [38], taking the raw wrist observation I_t and the encoder vector $\mathbf{e}_t$ as input and predicting an action chunk of length L consisting of a 6-DoF wrist action and the 20-DoF hand command. All policy training is performed on human-collected real data; no simulation, reinforcement learning, hand segmentation, or generative inpainting is involved.

Qualitative Study This qualitative comparison highlights why the data-collection interface matters for fine manipulation (Figure 6). During SEED-UMI collection, the operator directly performs AirPods insertion and can regulate delicate contacts through immediate visual and physical feedback. In contrast, teleoperation separates the operator from the object contact; the resulting replay can over-drive the motion and apply excessive force, which may damage fragile objects. This feedback advantage helps explain why SEED-UMI is better suited to precise contact-rich demonstrations than teleoperation.

5 Evaluation

5.1 Experiment Setup

Hardware. Our platform consists of a RealMan RX75 robot arm and a Wuji dexterous hand with four actuated DoFs per finger (20 in total). The human and robot wear the shared exoskeleton described in §3, preserving the outer contact geometry and camera placement across embodiments. A dorsal Intel RealSense T265 provides 6-DoF wrist pose, while a ventral fisheye camera with a 150° field of view supplies policy observations. Wrist trajectories use a fixed T265-to-tool-center-point (TCP) alignment without trajectory correction.

Policy	Screw Driving			AirPods Case			Ball Throwing			Table Cleaning			Air Freshener		
	T	P^-	P^+	T	P^-	P^+	T	P^-	P^+	T	P^-	P^+	T	P^-	P^+
ACT [36]	80.0	55.0	70.0	65.0	65.0	75.0	60.0	60.0	70.0	75.0	50.0	65.0	**90.0**	70.0	85.0
DP [37]	65.0	40.0	55.0	55.0	55.0	60.0	50.0	50.0	55.0	60.0	40.0	55.0	70.0	45.0	60.0
$\pi_{0.5}$ [38]	**90.0**	60.0	80.0	75.0	75.0	**85.0**	70.0	70.0	**80.0**	**85.0**	60.0	75.0	85.0	65.0	80.0

Table 1: **Main results.** Success rate (%) from 20 autonomous robot rollouts per entry. **T** = policies trained on robot-side teleoperation demonstrations; $\mathbf{P}^-$ = SEED-UMI without paired fine-tuning; $\mathbf{P}^+$ = SEED-UMI with paired fine-tuning.

Tasks. We evaluate five real-world tasks (Fig. 5) covering tool use, fine finger coordination, dynamic release, and sustained contact:

- **Screw Driving:** grasp an electric screwdriver, align the bit with a screw head, and maintain axial pressure and a stable grasp during driving.
- **AirPods Case Insertion:** pick up an AirPod and insert it into its charging case using index–middle finger coordination; the case magnets assist final alignment.
- **Ball Basket Throwing:** grasp a ball from the table, lift it, and release it into a target basket, requiring accurate grasp-to-release timing.
- **Table Cleaning:** extract a tissue, wipe the tabletop under sustained contact, and discard the tissue.
- **Air Freshener Spray:** securely grasp a spray can and initiate a downward actuator press through coordinated multi-finger support and thumb motion.

Comparison. We compare three conditions on the same robot platform, varying the demonstration source and encoder-to-command mapping:

- **Teleoperation (T).** The operator wears the exoskeleton, whose encoder readings are mapped directly to robot hand motor commands in real time. Policies are trained on demonstrations recorded on the robot side.
- **Ours w/o paired fine-tuning ($\mathbf{P}^-$).** Policies are trained on human-side demonstrations and deployed using the babbling-only mapping f_{θ^0}.
- **Ours w/ paired fine-tuning ($\mathbf{P}^+$).** Policies are trained on human-side demonstrations and deployed using the refined mapping f_θ from the full SEED-UMI pipeline (Fig. 3).

Training data and mapping calibration. Each policy is trained on 100 demonstration episodes per task, with initial object positions randomized within task-specific ranges. Paired fine-tuning uses 250 calibration episodes, with 50 episodes for each of five objects differing in size, shape, stiffness, and grasp type. For each paired calibration episode, the same object is rigidly fixed at one tabletop pose: the human grasps it, and the robot subsequently replays the motion through f_{θ^0}. The paired encoder traces supervise mapping refinement through Eq. 1; policy learning for P^- and P^+ uses only human-side demonstrations.

Evaluation protocol. We train ACT [36], Diffusion Policy (DP) [37], and $\pi_{0.5}$ [38] under each condition and evaluate each task–backbone–condition combination over 20 autonomous robot rollouts, also with randomized initial object positions. We report the percentage of rollouts that complete the task's main action sequence. For Air Freshener Spray, success requires a secure grasp and a clearly initiated downward press; full actuator depression and spray emission are not required because of the hand's actuation-force limit. The collection-efficiency comparison uses successful AirPods demonstrations collected by the same operator within a 30-minute window for each collection method, excluding the one-time paired-replay and optimization overhead.

5.2 Key Findings

F1: SEED-UMI improves fine and dynamic tasks. Averaged over all tasks and policy backbones, teleoperation reaches 71.7% success, while P^+ reaches 70.0% and P^- reaches 57.3%. Thus paired fine-tuning preserves nearly all of the robot-side data quality while removing the robot from the human collection loop. The task-level pattern is more informative: P^+ outperforms T on AirPods Case Insertion (73.3% vs. 65.0%) and Ball Basket Throwing (68.3% vs. 60.0%). These are precisely the settings where real-time teleoperation is least natural, because small orientation corrections and fast grasp-to-release timing are hard to execute through the delayed robot control loop.

F2: Paired fine-tuning matters most under contact load. Comparing P^- to P^+ isolates the contribution of paired-replay fine-tuning. The average lift is +12.7 percentage points, but it is concentrated on contact-load-dominated tasks: Screw Driving improves by +16.7 points, Table Cleaning by +15.0 points, and Air Freshener Spray by +15.0 points. This matches our expectation: the babbling-only mapping f_{θ^0} is trained in free space and drifts under finger compliance. The paired-replay loss (Eq. 1) corrects precisely this regime.

F3: SEED-UMI is significantly more efficient than teleoperation. We compare data collection throughput on the AirPods Case Insertion task within a fixed 30-minute window. The same operator collected 18 successful demonstrations via teleoperation and 52 via SEED-UMI, an almost $3.0\times$ throughput gain. Because P^+ and teleoperation have nearly matched overall rollout success (70.0% vs. 71.7%), the success-normalized gain in *useful training data per minute* remains approximately $2.9\times$ (Fig. 7). The practical benefit of SEED-UMI is therefore not that it sacrifices data quality for speed, but that it decouples human demonstration from real-time robot execution while paired fine-tuning recovers most of the quality gap.

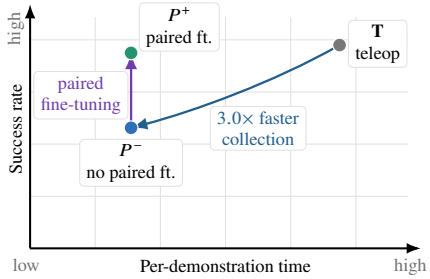

Figure 7: **Efficiency–quality trade-off.**

6 Conclusions

We present SEED-UMI, a scalable and efficient demonstration collection and policy learning framework that uses a shared exoskeleton as the interface between human demonstrations and robot execution. By letting the human operator and robot hand wear the same mechanism, SEED-UMI transfers natural contact-rich human motion into robot actions through paired encoder supervision while preserving aligned raw wrist observations. Through challenging real-world experiments, we demonstrate SEED-UMI's capability in learning precise, contact-rich, and dynamic dexterous manipulation policies, reaching 70.0% mean success while collecting data $3.0\times$ faster than teleoperation. Our work establishes a new approach to collecting real-world dexterous hand data efficiently and at scale beyond traditional teleoperation.

7 Limitations

Although SEED-UMI improves the efficiency of dexterous demonstration collection, several limitations remain. The current system is not yet fully automated across dexterous hands: adapting to different link layouts, joint limits, and actuation ranges still requires engineering choices in exoskeleton geometry, calibration motions, and paired regression targets. Because operators wear the exoskeleton during collection, wearability, donning time, and task convenience also affect long-horizon and delicate manipulation. Future work will automate cross-hand design and mapping, improve the wearing experience, and incorporate tactile sensing for more accurate contact estimation.

References

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[2] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

[3] 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.

[4] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[5] R. Meattini, R. Suarez, G. Palli, and C. Melchiorri. Human to robot hand motion mapping methods: Review and classification. *IEEE Transactions on Robotics*, 39(2):842–861, 2022.

[6] G. Du, P. Zhang, J. Mai, and Z. Li. Markerless kinect-based hand tracking for robot teleoperation. *International Journal of Advanced Robotic Systems*, 9(2):36, 2012.

[7] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[8] 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 (RSS)*, 2023.

[9] 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. *arXiv preprint arXiv:2403.07788*, 2024.

[10] F. L. Hammond, Y. Mengüç, and R. J. Wood. Toward a modular soft sensor-embedded glove for human hand motion and tactile pressure measurement. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[11] 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.

[12] 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. 2019.

[13] MANUS Meta. MANUS robotics gloves. https://www.manus-meta.com/robotics, 2025. Accessed 2026-05-06.

[14] P. Ben-Tzvi and Z. Ma. Sensing and force-feedback exoskeleton (safe) robotic glove. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(6):992–1002, 2014.

[15] H. Zhang, S. Hu, Z. Yuan, and H. Xu. DOGlove: Dexterous Manipulation with a Low-Cost Open-Source Haptic Force Feedback Glove. In *Robotics: Science and Systems (RSS)*, 2025.

[16] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[17] C. Xin, M. Yu, Y. Jiang, Z. Zhang, and X. Li. Analyzing key objectives in human-to-robot retargeting for dexterous manipulation. *IEEE Robotics and Automation Practice*, 2026.

[18] H. Liu, Z. Zhang, Z. Jiao, Z. Zhang, M. Li, C. Jiang, Y. Zhu, and S.-C. Zhu. A reconfigurable data glove for reconstructing physical and virtual grasps. *Engineering*, 32:202–216, 2024.

[19] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

[20] L. Y. Chen, K. Hari, K. Dharmarajan, C. Xu, Q. Vuong, and K. Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. *arXiv preprint arXiv:2402.19249*, 2024.

[21] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024.

[22] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik. Learning visuotactile skills with two multifingered hands. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[23] M. Zhu, A. Zhu, J. V. S. Ramos, B. J. Kim, Y. Shi, Y. Wu, R. Hou, Q. Wang, E. Song, T. Fan, et al. Whed: A wearable hand exoskeleton for natural, high-quality demonstration collection. *arXiv preprint arXiv:2602.17908*, 2026.

[24] H.-S. Fang, B. Romero, Y. Xie, A. Hu, B.-R. Huang, J. Alvarez, M. Kim, G. Margolis, K. Anbarasu, M. Tomizuka, et al. Dexop: A device for robotic transfer of dexterous human manipulation. *arXiv preprint arXiv:2509.04441*, 2025.

[25] J. Du, J. Ren, Q. Yu, N. Zhang, Y. Deng, X. Wei, Y. Liu, G. Gu, and X. Zhu. Mile: A mechanically isomorphic exoskeleton data collection system with fingertip visuotactile sensing for dexterous manipulation. *arXiv preprint arXiv:2512.00324*, 2025.

[26] X. Chen, Y. Pan, M. Li, and X. Ding. Dexvitac: Collecting human visuo-tactile-kinematic demonstrations for contact-rich dexterous manipulation. *arXiv preprint arXiv:2603.17851*, 2026.

[27] Z.-H. Yin, C. Wang, L. Pineda, F. Hogan, K. Bodduluri, A. Sharma, P. Lancaster, I. Prasad, M. Kalakrishnan, J. Malik, et al. Dexteritygen: Foundation controller for unprecedented dexterity. *arXiv preprint arXiv:2502.04307*, 2025.

[28] A. Gupta, C. Eppner, S. Levine, and P. Abbeel. Learning dexterous manipulation for a soft robotic hand from human demonstrations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[29] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[30] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[31] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. J. Fan, and Y. Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[32] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

[33] Z. Xue, S. Deng, Z. Chen, Y. Wang, Z. Yuan, and H. Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. *arXiv preprint arXiv:2502.16932*, 2025.

[34] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

[35] Y.-H. Wu, J. Wang, and X. Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on robot learning*, pages 618–629. PMLR, 2023.

[36] 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.

[37] 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*, 44(10-11):1684–1704, 2025.

[38] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025.