

DITTO: Dexterous Interface for Transparent TeleOperation

Joaquin Palacios^{*,1}, Katelyn Lee^{*,1}, Cheng Zhang², Zhanpeng He^{†,3}, and Matei Ciocarlie^{†,1}

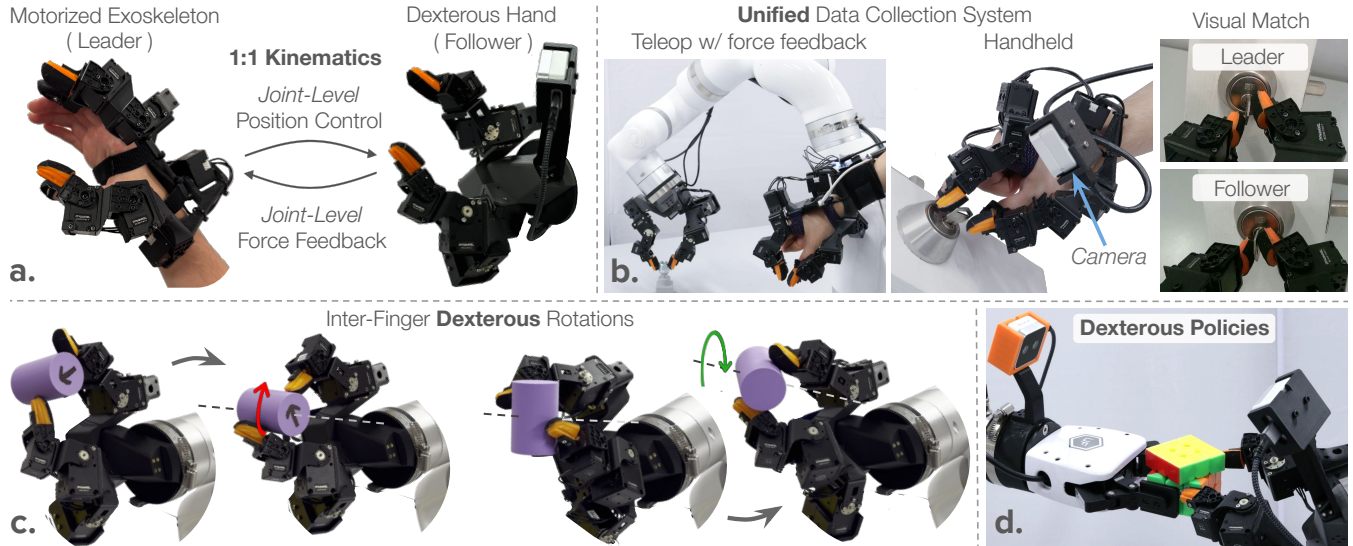


Fig. 1: **Overview of DITTO.** (a) 1-to-1 leader-follower design enables *joint-level* force feedback and allows for (b) a unified platform supporting both *bilateral teleoperation* and *handheld* (in-the-wild) data collection. DITTO’s anatomy-based design is capable of (d) human-like dexterous primitives, such as multi-axes rotations. We demonstrate dexterity and data quality via (e) contact-rich manipulation policies. Videos at: <https://roamlab.github.io/ditto/>.

Abstract—Collecting for manipulation with high-DOF hands is challenging, as interfaces must capture rich hand motion while rendering the contact interactions essential for precise manipulation. Existing data collection approaches face a trade-off: teleoperation ensures deployment consistency but lacks force feedback, while handheld (in-the-wild) systems provide natural force transparency but introduce a visual embodiment gap at deployment. We present DITTO, a Dexterous Interface for Transparent TeleOperation, which resolves this through the anatomically informed co-design of a dexterous 7-DOF robotic hand and a kinematically equivalent motorized exoskeleton. A 1-to-1 actuator mapping between the exoskeleton and robotic hand enables handheld (in-the-wild) data collection and bilateral teleoperation with *joint-level* force feedback unified in a single platform. We demonstrate that the DITTO exoskeleton spans the operator’s natural index-to-thumb workspace, and showcase DITTO’s dexterous capabilities via learned policies on contact-rich tasks. Videos are available at: <https://roamlab.github.io/ditto/>.

Index Terms—Teleoperation, Dexterous Manipulation, Exoskeleton

I. INTRODUCTION

Human demonstrations are a critical source of data for learning robot manipulation policies [1], [2]. In particular, two dominant paradigms have emerged for collecting manipulation demonstrations: *teleoperation* and *handheld (in-the-wild)*

collection. However, both paradigms remain challenging for highly articulated hands that use intrinsic DOFs for in-hand object reorientation and fine-grained regulation of contact forces between the hand, objects, and environment.

In *teleoperation*, human movement is recorded and used to command a remote robot manipulator to perform a task for data collection. This ensures that, from the perspective of the robot, demonstration data matches autonomous deployment. However, common teleoperation interfaces such as vision-based tracking [3] and sensor gloves [4] can make fine-grained control unintuitive due to lack of force feedback and the kinematic gap between the operator’s hand and the robot. Wearable motorized exoskeletons are an attractive alternative as they can provide force feedback at the fingertips [5], [6], but they do not resolve this kinematic gap.

In contrast, in the *handheld* or *in-the-wild* data collection paradigm, the demonstrator collects manipulation data directly through an interface physically attached to their hand [7]–[9]. This mode is scalable and intuitive: the demonstrator receives direct force feedback from physical contact, and the interface itself constrains motion to the target hand’s kinematics. However, it introduces a gap between data collection and deployment in both observation and action. For palm-mounted cameras, visual differences can be mitigated through in-painting [7] or hardware appearance matching [8], [9], but images from extrinsic cameras still include the operator, creating a visual mismatch. Moreover, because

* Equal contribution.

† Equal advising.

¹Dept. of Mechanical Engineering, ²Dept. of Computer Science, Columbia University. ³Dept. of Computer Science, Stanford University.

the robot’s actuators do not execute the demonstration, the collected actions do not faithfully capture contact-rich interactions. Lastly, although mechanically constraining the user to the target hand’s kinematics reduces the kinematic gap, if the robot’s kinematics are not well aligned with the user’s natural workspace, it can hinder their dexterity and natural motion.

These paradigms expose a trade-off: *teleoperation* collects high-fidelity, robot-native data, but high-DOF control is difficult due to a lack of kinematic and force transparency. *In-the-wild* interfaces provide intuitive collection with natural force feedback, but introduce mismatches between data collection and deployment in both observation and action.

We present DITTO, **Dexterous Interface for Transparent TeleOperation**, which bridges this trade-off through the anatomically informed co-design of a dexterous 7-DOF 2-finger (thumb–index) **robot hand** and a matching **motorized and wearable exoskeleton** (Figure 1). The key features of DITTO, and the main contributions of this paper, include:

- With only two fingers, DITTO uses a high-DOF kinematic structure to provide in-hand dexterity using intrinsic finger motion, without need for re-grasping or large wrist movement. In particular, our kinematic analysis on human manipulation data shows that DITTO’s workspace spans the natural human index-to-thumb dexterous workspace.
- DITTO provides a 1:1 kinematic and actuator correspondence between the exoskeleton and its corresponding robot hand. This co-design enables **joint-level force feedback** for force-aware teleoperation even for dexterous primitives such as continuous rotation and in-hand object reorientation.
- DITTO provides a unified platform supporting both scalable handheld data collection **and** bilateral teleoperation with force feedback, enabling complementary data sources within a single system. Our user study (n=8) shows that force-aware teleoperation produces faster and better demonstrations (lower unwanted forces), and that established methods such as Diffusion Policies can readily learn from both types of DITTO data across contact-rich tasks that leverage intrinsic dexterity.

II. RELATED WORK

Teleoperation for dexterous hands: Controlling high-DOF hands for dexterous tasks requires 1) fine-grained control and 2) regulating contact forces during manipulation. The former is challenging due to the kinematic gap between the operator’s and the robot’s hand; the latter is challenging due the lack of force feedback, relying purely on vision. Commonly used motion tracking-based systems [3] are lightweight and easy to use for simple grasping tasks, but are limited by these challenges in dexterous tasks.

Leader-follower designs like TILDE [10] can close the kinematic gap directly through joint-level correspondence. However, TILDE is not wearable, making arm control during teleoperation difficult, and lacks force feedback. A complementary line of work narrows the kinematic gap in software rather than hardware, augmenting kinematic

TABLE I: Comparison with previous devices.

Device	Tasks IHD	Teleoperation Force Feedback	In-The-Wild Supported	NIP	Observation Ex. Cam.
DOGlove [5]	✗	Single-Axis	✗	N/A	✓
GEX [6]	✗	Task-Space	✗	N/A	✓
DexUMI [7]	✓	✗	✓	✗	✗
DEXOP [8]	✓	✗	✓	✓	✗
DEX-Mouse [12]	✗	Single-Axis	✓	✓	✗
DITTO (ours)	✓	Joint-Level	✓	✓	✓

IHD = in-hand dexterity, or ability to reorient objects with intrinsic DOFs; NIP = no in-painting needed; Ex. Cam. = extrinsic camera without human operators in the frame.

retargeting with reinforcement-learned skill priors [4]; this can improve teleoperation fidelity but requires a trained skill prior and still provides no force transparency.

Motorized hand exoskeletons aim to improve force regulation by rendering force feedback to the operator during teleoperation [5], [6], [11]. However, in packaging actuators for force feedback, these devices forgo kinematic correspondence. This means they also rely on kinematic retargeting for motion. This limitation extends to force feedback, where devices either simplify feedback to a single-axis force per finger (i.e. DOGlove), or retarget forces in task space (i.e GEX [6] and N2D [11]), both of which do not provide understanding of which joint is applying/experiencing forces.

Handheld (In-the-wild) data collection: Handheld devices mitigate the issues of teleoperation by having users perform the task via a wearable data collection interface, directly feeling interaction forces during manipulation. DexUMI [7] is a wearable exoskeleton that constrains the user’s hand to the target kinematics, but requires in-painting to reduce the visual mismatch compared to the target hand. DEXOP [8] and DexEXO [9] instead use linkages connecting the user’s fingertips to the target hand, positioning the palm camera to capture the actual robot and reduce the visual mismatch (in the palm camera), eliminating the need for in-painting.

However, embodiment gaps remain in two forms: 1) external camera views still include the operator, creating a visual mismatch (not present for teleoperation), and 2) the user performs the actions and applies forces, so demonstrations do not capture the robot’s actuator dynamics. DEX-Mouse [12] is a hybrid teleoperation and handheld system where the user wears both the target hand and a motorized wearable, controlling the hand’s wrist pose with their own arm and teleoperating the robot’s finger motion to fully capture the hand’s actuator dynamics. However, this mode still does not support extrinsic camera views without visual mismatch, and is limited in dexterity due to one DOF and actuator per finger.

From a kinematic standpoint, this family of devices constrain the user’s hand movements to the target hand’s kinematics, which can constrain the human hand’s natural workspace. DEXOP [8] mitigates the workspace constraint by co-designing the wearable interface and robotic hand, but does not provide quantitative kinematic analysis of how well the workspace aligns with human hand.

Table I places DITTO in the context of existing devices. DITTO co-designs a wearable, *motorized* exoskeleton with a matching 7-DOF thumb–index follower hand. The resulting 1:1 actuator mapping enables joint-level force feedback during bilateral teleoperation, supports handheld in-the-wild

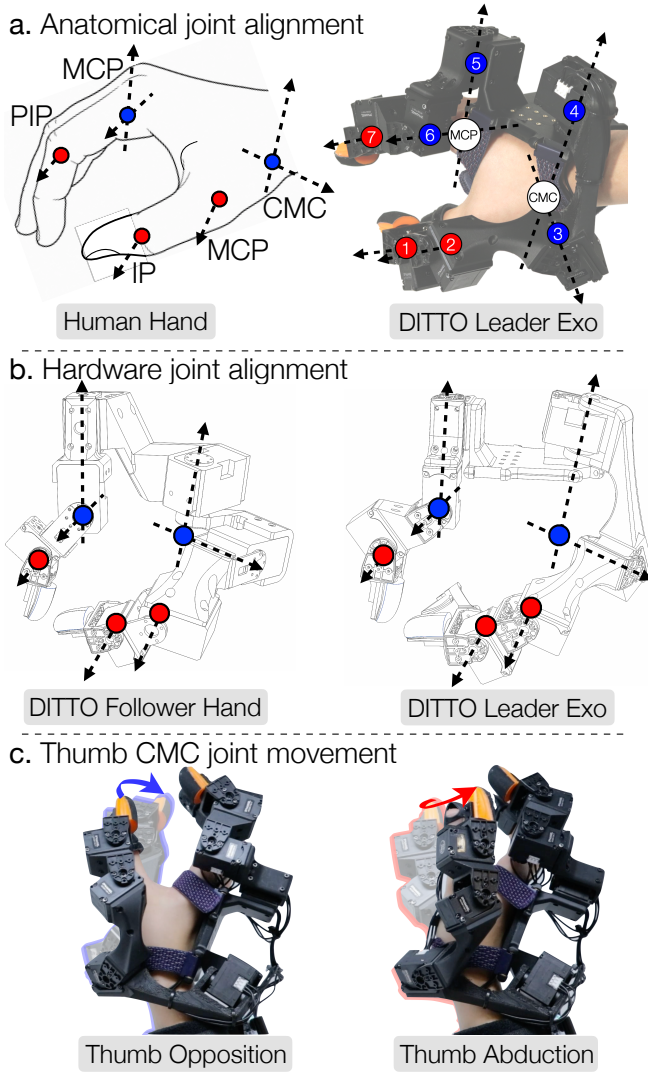


Fig. 2: (a) DITTO aligns each of its 7 motor axes with a finger joint axis. In blue, the 2-DoF CMC and MCP joints have offset, decoupled motors so the motor axes intersect at the anatomical joint center. (b) DITTO has 1:1 kinematic equivalence between the leader exoskeleton and the follower hand. (c) The offset, decoupled motors for the CMC enable dexterous thumb motions like abduction and opposition.

collection without in-painting, and, in teleop mode, allows extrinsic cameras without the operator in frame, while retaining in-hand dexterity via intrinsic finger motion.

III. DITTO HARDWARE DESIGN

DITTO aims to bridge the fundamental trade-off between kinematic and force transparency in teleoperation by co-designing a dexterous 7-DOF hand and a kinematically equivalent motorized wearable exoskeleton. This leader-follower system has a 1:1 actuator mapping, where every motor on the follower hand corresponds to a motor on the leader exoskeleton. This concept has been demonstrated for arm teleoperation [13] but not for high-DOF hands in a wearable form factor. The primary difficulty lies in the non-

obtrusive, wearable packaging of enough actuators to achieve complex tip-to-tip movement without restricting the user’s natural workspace. In this section, we describe how we achieve this via careful, anatomically-inspired kinematic co-design and how the 1:1 actuator mapping enables joint-level force rendering and transparency during teleoperation.

A. DITTO Kinematics

a) *Leader Kinematics:* We designed DITTO around the human hand, using anatomical finger kinematics as the basis for both the wearable exoskeleton (leader) and the robotic hand (follower). For the exoskeleton, we prioritize wearability and anatomical joint alignment, ensuring the exoskeleton axes coincide with the biological joint axes of the index finger and thumb. The DITTO index finger has three DOF: one aligning with the PIP (proximal interphalangeal) joint, and two aligning with the MCP (metacarpophalangeal) joint. The DITTO thumb has four DOF: two aligning with the IP (interphalangeal) and MCP flexion joints, and two aligning with the CMC (carpometacarpal) joint. The CMC joint is a complex biological joint whose motion is commonly approximated as a two DOF universal joint with perpendicular axes, and we adopt this simplification in our design [14].

To accommodate the servo motors within the form factor of a wearable hand exoskeleton, we position them along the plane formed by the tip-to-tip pinch grasp between the index finger and thumb (Fig. 2a.). For the index finger PIP joint and the thumb flexion degrees of freedom, this placement is straightforward, as the joint axes align naturally with motor placement in this plane. For the index MCP and thumb CMC joints—each of which we model as a two DOF universal joint—direct motor placement is more constrained. We address this with a non-anthropomorphic design that decouples each universal joint into two offset joints whose axes intersect at the anatomical center of rotation of the human joint (Fig. 2c.). This decoupled arrangement allows each motor to align with one rotational axis of the universal joint while remaining positioned around the hand.

The exoskeleton attaches to the operator’s hand at three points of contact: a rigid handpiece on the dorsal side of the hand and two finger caps on the distal links of the index finger and thumb. Users don the exoskeleton by aligning with their index MCP joint and securing the base to the dorsal side of their hand. Each fingertip consists of a silicone cap and a rigid piece that connects the user’s fingertips: one attaches at the PIP joint on the index finger frame, and the other at the IP joint on the thumb frame. To accommodate a range of hand sizes, these finger interface pieces can be customized to different finger lengths.

b) *Follower Kinematics:* We derive the robot follower hand kinematics directly from the exoskeleton leader to keep the same joint positions and axes for 1:1 kinematic transparency. Without the space constraint of the operator’s hand on the exoskeleton leader, we can use larger servo motors with greater output torque for the robot follower hand. The larger motors impose space constraints and require two joints to be translated along the joint axis: the index MCP

abduction joint by 12 mm and the first thumb CMC joint by 33.1 mm. All remaining links are kinematically identical to the leader, preserving joint-level transparency between the operator and the follower hand.

B. Hardware characteristics

The leader exoskeleton is low-cost at $\approx \$900$ and weighs 430 g; donning and doffing take about 20 s and 10 s, respectively. The follower hand weighs 935 g and costs $\approx \$1500$, a substantially lower price than typical commercial dexterous hands. We selected Dynamixel XC330-T181 motors for the leader exoskeleton for their small size and because their 181:1 gear ratio was small enough to allow users to comfortably backdrive them during motion capture. The follower hand uses Dynamixel XM430 motors for all joints except the most distal index PIP and thumb IP joints, which use smaller Dynamixel XC330-T288 motors.

C. Teleoperation with Force Feedback

DITTO's leader/follower 1:1 actuator mapping enables joint-level force feedback with direct per-joint correspondence. This has the added benefit of reducing bilateral teleoperation to simple per-joint loops:

- **Leader to follower:** Joint-level position tracking via Dynamixel position control.
- **Follower to leader:** Force feedback on the leader is computed independently for each joint using the corresponding follower motor current. A deadband is applied to suppress feedback during free-space motion:

$$\tau_l = \begin{cases} -k_f (\tau_f - \tau_{th}), & \text{if } |\tau_f| > \tau_{th} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where τ_f denotes the estimated torque from motor current of the follower actuator, k_f is a feedback gain, and τ_{th} defines a deadband threshold that filters out low-magnitude currents associated with general motion, while capturing high-magnitude currents from contacts and applied forces.

This structure eliminates the need for task-space force mapping and reduces computational overhead. Our system supports bilateral teleoperation at 200 Hz using off-the-shelf U2D2's for communication with Dynamixel motors. User safety is enforced via conservative joint torque limits (200 mA $\approx$ 0.08 Nm), position limits, and an emergency stop.

D. Handheld Data Collection

While DITTO can be used for leader-follower teleoperation, the DITTO exoskeleton by itself can also be used directly for handheld data collection. To this end, the DITTO leader is equipped with a camera and a pose tracking sensor. The camera positioning matches that of camera on the follower hand and is carefully placed to ensure only the exoskeleton is visible and not the user's hand, ensuring minimal visual gap without requiring post-processing or in-painting (Fig. 1.b).

IV. USER STUDY: DEXTEROUS DATA COLLECTION

When used in leader-follower teleoperation mode, DITTO aims to provide active force feedback while still allowing rich tip-to-tip movement for dexterous tasks. To assess the importance of these features, we conducted a user study ($n = 8$) comparing DITTO teleoperation against two ablations and across two tasks that require coordinated fingertip dexterity and force regulation. We compare three **teleoperation conditions**:

- **DITTO** exoskeleton: leader-follower teleoperation using DITTO exoskeleton with force feedback enabled.
- **DITTO-NF** (No Force): an ablation of DITTO's bilateral force rendering. The demonstrator uses the DITTO leader exoskeleton joint position commands are related to the follower hand, but force feedback on the leader is disabled.
- **Fingertip**: an ablation of DITTO's joint-level kinematic correspondence. The leader exoskeleton is not used at all; instead, we attach 6-DOF pose trackers to the operator's fingertips and retarget fingertip motion to the DITTO follower hand via inverse kinematics.

For each teleoperation condition, subjects completed two **tasks**: *mahjong tile* and *ratchet*. In the *mahjong tile* task, subjects flip a mahjong tile as many times as possible within a 60-second trial. The tile rests on a platform mounted to a 6-axis force-torque sensor, and subjects teleoperate a 7-DOF robot arm while using the DITTO follower hand to manipulate the tile. A successful flip is defined as a clockwise, 90 degree rotation about the tile's long axis. This task evaluates speed in fingertip dexterity, as successful manipulation requires coordinated index-thumb motion to quickly grasp and rotate the tile. In the *ratchet* task, subjects rotate a ratchet for three full revolutions as fast as possible. The 7-DOF robot arm is locked into a fixed pose relative to the ratchet to isolate interaction forces to the DITTO follower hand to evaluate force regulation and whether bilateral force rendering reduces excessive spurious forces during manipulation.

Protocol All $n = 8$ subjects were recruited via our IRB-approved protocol and had no prior experience with hand teleoperation. To mitigate order effects across teleoperation conditions and minimize exoskeleton donning and doffing time, we used a hierarchical randomization procedure via coin flip. For each participant, we first randomized whether trials began with **Fingertip** tracking or a DITTO-based condition. The order of the two DITTO-based conditions, **DITTO-NF** and **DITTO**, was then randomized independently. The order of tasks (*mahjong tile*, BOA) was kept constant to ensure that any task learning effects were consistent across participants and conditions.

Metrics We evaluated performance using four metrics [Table II]. The total flips (#) is the throughput of successful flips completed within the 60-second trial on the *mahjong tile* task, aggregated over three trials per condition, capturing the manipulation efficiency of each teleoperation condition. For the *ratchet* task, we measure task completion time, defined as the total time required

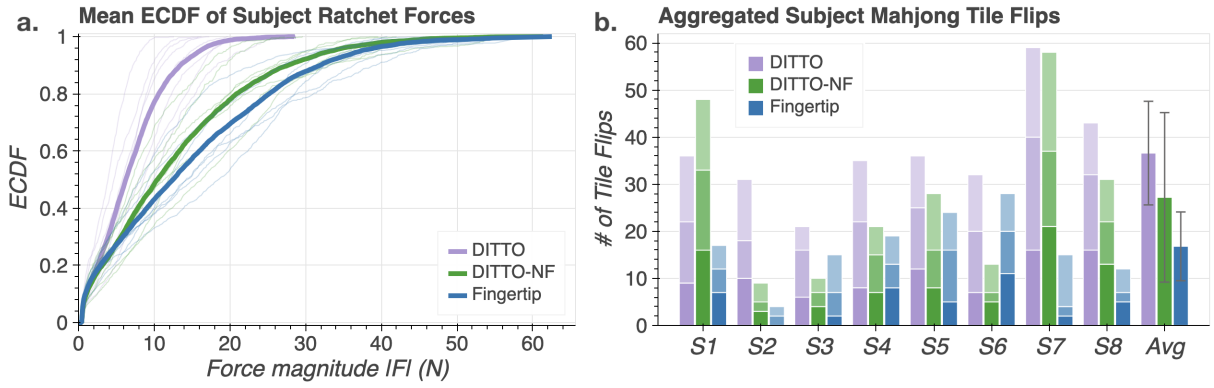


Fig. 3: (a) The three bolded lines in the ECDF plot denote the gross average over all $n = 8$ subjects per teleop condition. Remaining lines denote the individual subject curves. (b) The stacked bar plot tallies the total number of mahjong tile flips over the three trials with teleop condition labeled by color and trial number denoted by opacity.

TABLE II: User study metrics (avg & std for 8 users).

task	metric	DITTO	DITTO-NF	Fingertip
mahjong tile				
flips (#)		36.6 ± 11.0	27.2 ± 18.0	16.8 ± 7.3
ratchet				
time (s)		24.1 ± 6.6	22.4 ± 9.1	34.2 ± 8.1
force (N)		6.9 ± 2.0	13.0 ± 3.0	14.8 ± 2.5
% ≤ 14.8N		92.8 ± 8.7	64.8 ± 11.5	56.7 ± 10.7

to complete three full revolutions. We also measure the interaction force magnitude with a 6-axis f/t sensor mounted beneath the ratchet; lower values indicate finer force control.

To characterize the force distribution during active manipulation in the ratchet condition, we aggregated all trials per condition and computed the empirical distribution function (ECDF), $F(x) = \frac{\# \text{ of points } \leq x}{n}$, where x is the force magnitude, and n is the total number of samples in the condition. We can define a reference threshold at 14.8 N, the average force value from the Fingertip condition [Table II], and calculate the ECDF value for each condition, denoted as % ≤ 14.8N.

Statistical significance One-sided permutation tests found statistically significant differences for DITTO vs. Fingertip across all metrics (all $p \leq 0.023$) and for DITTO vs. DITTO-NF for the two ratchet force metrics ($p = 0.012$). ratchet times for DITTO-NF are slightly faster than DITTO, but elevated ratchet forces suggest that users, without force feedback, speed up task completion with more aggressive behavior.

Mahjong Tile results Aggregating across all three trials, every subject completed the highest number of successful flips with the DITTO exoskeleton, and all but S1 achieved their highest total with force rendering enabled [Fig. 3a.]. On average, subjects completed 36.6 ± 11.0 total flips using DITTO, compared to 27.28 ± 18.0 with DITTO-NF and more than $2\times$ the 16.8 ± 7.3 total flips under Fingertip II. The gap between Fingertip and the two exoskeleton conditions indicates that DITTO’s kinematic structure supports more effective coordinated fingertip manipulation than

fingertip retargeting alone.

We defined two failure cases for the mahjong tile task: the tile getting flicked off of the platform and overloading of the force-torque sensor due to excessive contact forces. These failures occurred in 6/8 subjects, and only occurred under the Fingertip and DITTO-NF conditions. For S2, all Fingertip trials and 2 of 3 DITTO-NF trials failed due to sensor overloading (force values of ≥ 100 N). These failure cases suggest that force feedback also helps subjects regulate object interaction forces during manipulation, preventing excessive forces that could damage the manipulated object or exceed sensor limits.

Ratchet results For the ratchet task, DITTO yielded the lowest average interaction forces for every subject at 6.9 ± 2.0 N, compared to 13.0 ± 3.0 N under DITTO-NF and 14.8 ± 2.5 N under Fingertip [Table II]. We filter out force magnitudes below 0.45 N ($3\times$ the 0.15N resolution of the force-torque sensor) to exclude idle periods between active manipulation, which can bias the average force downward and obscure differences in force regulation during the task.

The ECDF plots show that DITTO samples are concentrated at lower forces compared to the other teleoperation conditions, with the DITTO lines having a steeper and shorter curve than DITTO-NF and Fingertip lines [Fig 3]. Notably, Table II shows that for all subjects, DITTO has the largest percentage of force values below the threshold, indicating that force feedback reduces the prevalence of high-force interactions.

While DITTO-NF achieved the fastest average completion times, these trials also exhibited higher interaction forces, suggesting that without force feedback, operators applied more aggressive forces to complete the task quickly. This indicates that the higher forces under DITTO-NF and Fingertip were *not* essential for efficient performance, but rather a consequence of *reduced force awareness* during teleoperation.

V. KINEMATIC WORKSPACE ANALYSIS

Matching human index-to-thumb dexterity is a key design goal for DITTO providing two important benefits. First,

	Ratchet		Cube		Cap		USB		Mahjong		Average	
System	Succ.	X-Err	Succ.	X-Err	Succ.	X-Err	Succ.	X-Err	Succ.	X-Err	Succ.	X-Err
DITTO	1.0±0.0	4.2±1.6	1.0±0.1	4.0±2.2	1.0±0.0	3.0±0.9	0.8±0.2	1.7±0.7	1.0±0.0	3.4±1.1	0.96±0.13	3.26±1.67
DexUMI	0.1±0.2	32.1±8.1	0.3±0.3	28.2±9.3	0.6±0.4	17.8±7.9	0.1±0.1	31.2±10.7	0.2±0.3	30.6±12.5	0.26±0.34	27.98±11.16
Doglove	0.5±0.2	12.1±5.4	0.5±0.3	13.5±7.5	0.3±0.3	18.4±7.3	0.1±0.2	21.2±6.1	0.2±0.3	18.6±6.0	0.32±0.31	16.76±7.35

TABLE III: Kinematic workspace coverage across five manipulation tasks. **Succ.** denotes the fraction of sampled poses satisfying the 1 mm position tolerance and orientation check. The orientation check requires the fingertip x-axis to lie within 20° of the target axis and remain in the correct yz-hemisphere. **X-Err** is the mean x-axis angular error. **Average** reports the aggregate success rate and mean x-axis angular error across all tasks. Values are mean ± std over all trials and participants ($n = 7$). **Cube** averages four rotation directions; **Cap** averages cap and uncapped trials.

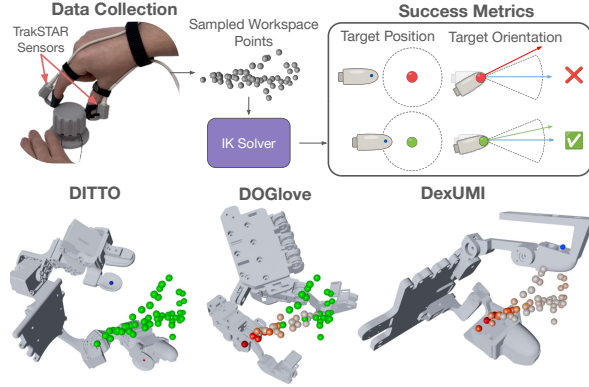


Fig. 4: Top: Kinematic analysis data collection and data processing workflow. Bottom: Example workspace results visualized, targets colored by IK solution quality, from green (within tolerance) to red (error exceeds threshold).

from the perspective of the DITTO leader (a wearable exoskeleton), a human-like workspace removes constraints on the teleoperator’s hand to allow natural and intuitive movement without pressure points. Second, from a follower perspective, human-like tip-to-tip movement implies high dexterity, since the human hand can perform complex manipulation with just index-to-thumb grasps exclusively.

To assess the DITTO tip-to-tip workspace and compare it against the human hand, we performed a kinematic analysis with data collected from seven subjects of varying hand sizes. We quantitatively compare the DITTO leader exoskeleton both against the human thumb-to-index workspace and against two other high-DOF wearable exoskeletons: DOGlove [5] and DexUMI-Xhand exoskeleton [7].

Data collection: We attach two 6-DOF motion capture sensors (trakSTAR, Ascension) to the index and thumb fingertips and record subjects performing several tasks that require coordinated index-thumb manipulation, capturing their relative range of motion during each task. Specifically, subjects completed eight tasks across three trials each: mahjong tile rotation, USB flash drive extrusion, ratchet rotation, test tube capping and uncapping, and clockwise/counterclockwise rotation of the top and right faces of a Rubik’s cube. Because each subject’s kinematic data defines the target poses, the reachable workspace depends directly on hand size. Across our seven subjects, index finger length had a range of 27 mm, palm width 31 mm, and thumb length 19 mm.

Kinematic analysis: After collecting the human kinematic data, we downsample the workspace using SE(3) Farthest Point Sampling (FPS) to get a subset of poses with near-uniform coverage of the original human workspace. Then, to assess relative pose of the index finger with respect to the thumb, we calculate the relative transform from the thumb tip to the index tip [Fig. 4].

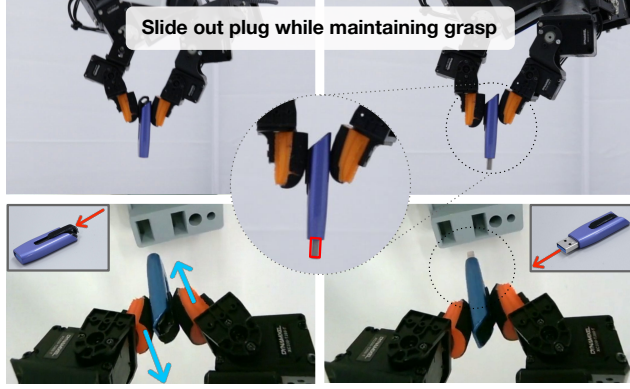
For a given exoskeleton design, we are interested if the robot kinematics can achieve the same tip-to-tip relative poses as observed in the human data set. Thus, we take the kinematic chain of the exoskeleton, set the thumb tip as a fixed base, then compute inverse kinematics using mink IK [15] for the index tip to check if a specific relative pose is feasible within a 1 mm position tolerance and 20° angular tolerance of the fingertip midline axis (defined as the x-axis). We then aggregate these results over the totality of the human tip-to-tip workspace and assess a robot’s kinematic capability based on what percentage of human tip-to-tip poses it is capable of reproducing with low position and angular error.

Results for all three exoskeletons are shown in Table III. DITTO reaches nearly the entire human thumb-to-index workspace, with a mean success rate at or near 1.0 and a mean x-axis orientation error of only 1.7° to 4.2°. Within the same 1 mm position and 20° orientation bounds, DexUMI and DOGlove reach at most 0.6 and 0.5, respectively, with orientation errors 3× to 8× larger. We note that all systems achieve near perfect *position-only* success rates across all tasks (exceptions being DexUMI reaching 0.4 ± 0.1 on USB and 0.8 ± 0.1 on mahjong, and DITTO reaching 0.9 ± 0.1 on USB). The drop from perfect position accuracy due to the 20° orientation constraint indicates this gap arises almost entirely from orientation: the anatomically aligned joint axes of the DITTO leader allow it to match both the position and orientation of natural fingertip poses. We believe this plays an important role in the fact that DITTO enables fast and high-quality data collection even for tasks that leverage intrinsic DOFs of the hand.

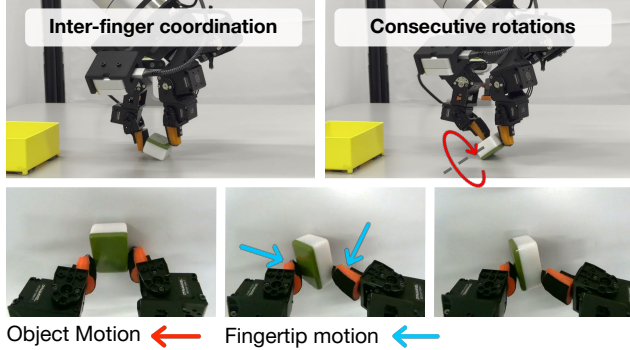
VI. POLICY LEARNING RESULTS

Finally, we evaluate whether DITTO data-collection enables autonomous policies for tasks requiring use of the intrinsic hand DOFs, as well as regulation of contact forces with the environment. We test these abilities for both data collection modes offered by DITTO, teleoperation and handheld, as well as combined datasets from both paradigms.

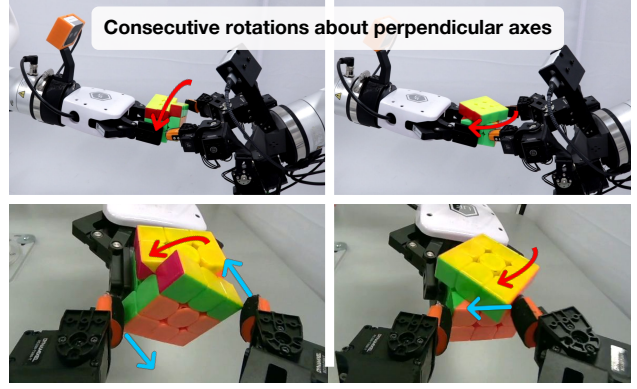
USB Drive: Controlled sliding



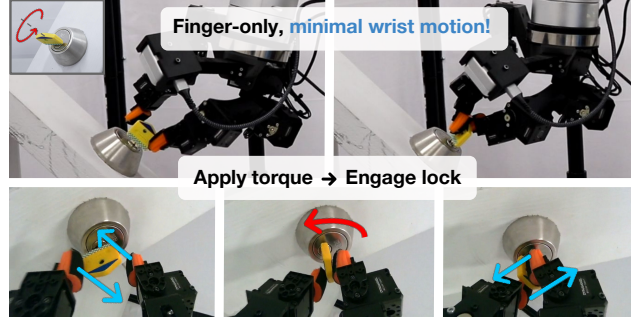
Mahjong Tile: Inter-finger reorientation



Rubik's Cube: Multi-Axes Rotation



Key: Continuous rotation



Object Motion ← Red arrow Fingertip motion ← Blue arrow

Fig. 5: Learned policies for dexterous and contact-rich manipulation tasks. For unimanual tasks, we include: **USB Drive Insertion**, which requires picking up and extending a retractable USB flash drive and then inserting it into a slot, **Mahjong Tile Flipping**, where the robot reorients a tile through two sequential flips followed by placement into a target box, and **Key Rotation** to engage a lock. For the bimanual task, we include **Rubik's Cube Rotation**, which requires picking up and then rotating two faces of a Rubik's Cube using thumb-index finger manipulation. Videos at: <https://roamlab.github.io/ditto/>

Unimanual Tasks	
Stage	Success
Key Rotation (60 Demos, 180 Epochs)	
Rotate key until lock engages	20/20
Mahjong Flipping (56 Demos, 250 Epochs)	
Rotate twice	20/20
Pick and place	20/20
USB Insertion (55 Demos, 300 Epochs)	
Pick up USB	20/20
Extend USB	19/20
Insert into slot	14/20
Bimanual Tasks	
Stage	Success
Rubik's Cube (59 Demos, 300 Epochs)	
Rotate right face	18/20
Rotate top face	17/20

TABLE IV: Task success rate for policies trained on teleoperation data collected using DITTO.

For the autonomous policies, we use standard Diffusion Policies with a Transformer Architecture [16] using the relative pose representation from [17].

A. Policies Trained with Teleoperation Datasets

We evaluate manipulation performance across unimanual and bimanual dexterous tasks [Fig. 5]. Our results in Table IV show that policies trained on DITTO-collected data achieve

high success rates. Notably, USB insertion requires both in-hand manipulation and extrinsic contact reasoning, showing that DITTO can capture nuanced finger coordination for complex interactions. On bimanual tasks, our policies also achieve high success rates, highlighting DITTO's ability to support coordinated two-hand manipulation and collect high-quality demonstrations beyond the unimanual setting.

B. Policies Trained with Handheld and Combined Datasets

To validate our handheld data collection capability, we evaluate and compare policies trained on *handheld* data, *teleoperated* data, and various mixtures of the two. We focus on a unimanual task requiring high finger-to-finger coordination: rotating a test tube cap to loosen it, then grasping and removing it without dropping it.

Our results show that for the rotate-and-remove subtask, the policy trained on *handheld* data alone performs similarly to those mixed with *teleoperation* data (small 5% margin of difference), showing that *handheld* data captures the finger coordination needed for continuously rotating a cap. For the retain-grasp-and-lift subtask, however, policies trained on *handheld* data alone fail entirely. The primary failure mode is a poor grasp before lifting, which leads to dropping the object. Augmenting with a small amount of teleoperation data (10 episodes) raises success to 70%, consistent with prior

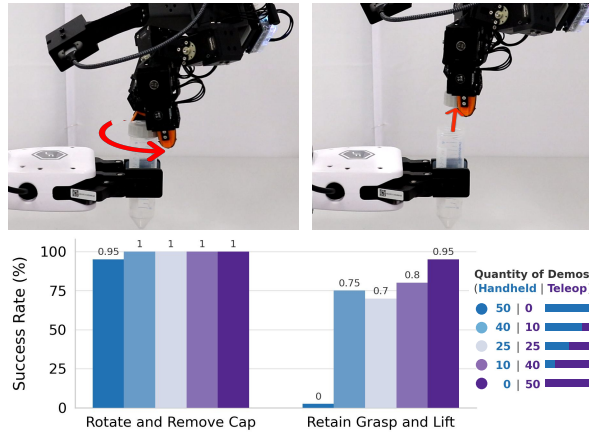


Fig. 6: Test tube uncapping policy performance for data mixing combinations of *handheld* and *teleoperation* demonstrations. All policies trained with 50 demos for 250 epochs.

work claiming handheld data alone is insufficient and that some *teleoperated* demonstrations are needed [8]. Training on teleoperation data alone further improves success by 20 percentage points. We hypothesize that the *handheld*-only failure reflects an action mismatch: applying forces and movement is performed by the user rather than the robot’s actuators, whereas teleoperated demonstrations record robot-native contact interactions. Overall, a small teleoperation fraction recovers performance relative to *handheld* alone, but teleoperation remains the stronger data source for this task.

VII. CONCLUSIONS AND LIMITATIONS

In this work, we introduced DITTO, an active 7-DOF wearable thumb-to-index exoskeleton and a robot hand with identical kinematic structure. The DITTO exo can be used both in handheld data collection and leader-follower teleoperation with joint-level force feedback. For both paradigms, the data collected allows training of autonomous manipulation policies executed on the follower robot. The user study, kinematic analysis, and policy learning results presented here support the following conclusions:

- DITTO matches human index-to-thumb tip-to-tip dexterity very well (96% of poses used by seven human subjects over five tasks), and better than comparable devices. A human-like workspace is desirable both as a measure of dexterity and to mitigate constraining an operator’s hand during data collection.
- Joint-level force feedback, enabled by the 1:1 kinematic mapping between the leader and the follower, improves both the speed and quality of teleoperated demonstrations.
- Data collected by DITTO allows training of effective autonomous policies that leverage intrinsic DOFs for complex tasks (flipping, inserting, manipulating an object’s internal DOFs), without regrasping or large wrist motions.
- DITTO allows both teleoperated and handheld data collection, with force feedback in both cases, and these two data streams can be combined for efficient policy training.

The key limitations of this work span hardware and data collection. Currently, DITTO is a high-DOF two-finger device.

Given the impressive capabilities of the human hand even when using only a thumb-to-index grasp, we believe that this is an avenue worthy of investigation. Nevertheless, the absence of additional fingers and a palmar surface limits the types of manipulation we can engage in. A natural extension would incorporate these components, enabling richer grasp taxonomies. Additionally, DITTO lacks tactile sensing, which is ultimately necessary for policies to make accurate contact decisions. Integrating tactile sensors remains an important direction for future work.

On the data collection side, how to best leverage the complementary nature of our two modalities remains an open question. A promising direction is to use in-the-wild collection for large-scale pretraining and teleoperation for high-quality behavior alignment. While our experiments show early evidence of the benefit of mixing these sources (Fig. 6), understanding co-training requirements at scale, particularly for VLA Models, is an important avenue for future work.

ACKNOWLEDGMENT

We disclose the use of generative AI for figures and code.

REFERENCES

- [1] Physical Intelligence et.al, “ $\pi_{0.5}$: a vision-language-action model with open-world generalization,” *arXiv preprint arXiv:2504.16054*, 2025.
- [2] J. Bjorck et.al, “Gr00t n1: An open foundation model for generalist humanoid robots,” *arXiv preprint arXiv:2503.14734*, 2025.
- [3] Y. Qin et.al, “Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system,” *arXiv preprint arXiv:2307.04577*, 2023.
- [4] Z.-H. Yin et.al, “Dexteritygen: Foundation controller for unprecedented dexterity,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.04307>
- [5] H. Zhang, S. Hu, Z. Yuan, and H. Xu, “DOGlove: Dexterous Manipulation with a Low-Cost Open-Source Haptic Force Feedback Glove,” in *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025.
- [6] Y. Dong, X. Liu, J. Wan, and Z. Deng, “Gex: Democratizing dexterity with fully-actuated dexterous hand and exoskeleton glove,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.04982>
- [7] 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.
- [8] H.-S. Fang et.al, “Dexop: A device for robotic transfer of dexterous human manipulation,” *arXiv preprint arXiv:2509.04441*, 2025.
- [9] A. Zhu et.al, “Dexexo: A wearability-first dexterous exoskeleton for operator-agnostic demonstration and learning,” 2026. [Online]. Available: <https://arxiv.org/abs/2603.17323>
- [10] Z. Si, K. L. Zhang, Z. Temel, and O. Kroemer, “Tilde: Teleoperation for dexterous in-hand manipulation learning with a deltahand,” in *Proceedings of (RSS) Robotics Science and Systems*, March 2024.
- [11] K. Calimbahin, J. Honma, Y.-T. Huang, L. Li, O. Hernandez, and M. C. Yip, “The N2D haptic glove: A multi-finger glove for 2D directional force feedback for contact rich manipulation.” [Online]. Available: https://kailin-calimbahin.com/papers/Haptic_Glove_Paper.pdf
- [12] 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. [Online]. Available: <https://arxiv.org/abs/2604.15013>
- [13] T. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware.”
- [14] A. L. Ladd et.al, “The thumb carpometacarpal joint: anatomy, hormones, and biomechanics,” vol. 62, pp. 165–179.
- [15] K. Zakka, “Mink: Python inverse kinematics based on MuJoCo,” Feb. 2026. [Online]. Available: <https://github.com/kevinzakka/mink>
- [16] C. Chi et.al, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.
- [17] C. Chi et. al, “Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots,” in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.