DexUMI: Using Human Hand as the Universal Manipulation Interface for Dexterous Manipulation

Mengda Xu [*,1,2,3] **Han Zhang** [*,1] **Yifan Hou** [1] **Zhenjia Xu** [5]
Linxi Fan [5] **Manuela Veloso** [3,4] **Shuran Song** [1,2]
[1] Stanford University, [2] Columbia University,
[3] J.P. Morgan AI Research, [4] Carnegie Mellon University, [5] NVIDIA

***Abstract*—We present DexUMI - a data collection and policy learning framework that uses the human hand as the natural interface to transfer dexterous manipulation skills to various robot hands. DexUMI includes hardware and software adaptations to minimize the embodiment gap between the human hand and various robot hands. The hardware adaptation bridges the kinematics gap using a wearable hand exoskeleton. It allows direct haptic feedback in manipulation data collection and adapts human motion to feasible robot hand motion. The software adaptation bridges the visual gap by replacing the human hand in video data with high-fidelity robot hand inpainting. We demonstrate DexUMI's capabilities through comprehensive real-world experiments on two different dexterous robot hand hardware platforms, achieving an average task success rate of 86%.**

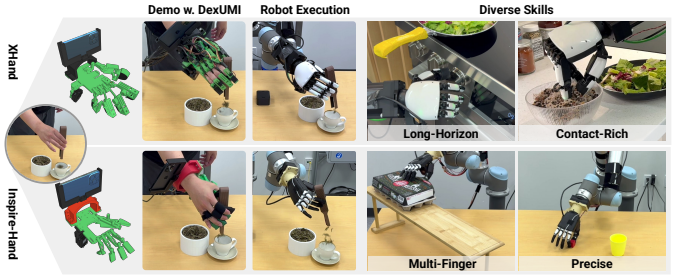

Fig. 1: **DexUMI** transfer dexterous human manipulation skills to various robot hand by using wearable exoskeletons and a data processing framework. We demonstrate DexUMI's capability and effectiveness on both underactuated (e.g., Inspire) and fully-actuated (e.g., XHand) robot hand for a wide variety of manipulation tasks.

I. Introduction

Human hands are incredibly dexterous in a wide range of tasks. Dexterous robot hands are designed with the hope of replicating this capability. However, it remains a significant challenge to transfer skills from human hands to robotic counterparts due to their substantial *embodiment gap*. This gap manifests in various forms, such as differences in kinematic structures, contact surface shape, available tactile information, and visual appearance.

What further complicates this challenge is the diversity of dexterous hand hardware designs available today. Each robotic hand presents different engineering trade-offs in degrees of freedom, motor ranges, actuation mechanisms, and overall dimensions. The solution for reducing the embodiment gap must handle the vast hardware design space. Teleoperation has become a popular manipulation interface for dexterous hands. However, teleoperation can be difficult due to the spatial observation mismatch and the lack of direct haptic feedback. These problems do not exist when human hand can perform the manipulation task directly. In other words, human hand itself is a better manipulation interface. In this paper, we ask the following question: How can we minimize the embodiment gap, so that we can use the human hand as the universal manipulation interface for diverse robot hands? To answer this question, we propose **DexUMI**, a framework with hardware and software adaptation components that is designed to minimize the action and observation gaps.

The **hardware adaptation** takes the form of a wearable hand exoskeleton. A user can directly collect manipulation data while wearing it. The exoskeleton is designed for each target robot hand through *a hardware optimization framework* that refines exoskeleton parameters (e.g., link lengths) to closely match the robot finger trajectories while maintaining wearability for the human hand. The hardware adaption provides the following benefits:

- **Intuitive demonstration with direct haptic feedback:** Unlike teleoperation systems, the wearable exoskeleton has no spatial mismatch and allows users to directly contact objects during manipulation, making the demonstration intuitive and doable without a robot.
- **Records feasible motion for the robot hand:** The exoskeleton constrains human hand motions to match the kinematics of the target hand, ensuring the recorded motion is transferable.
- **Capturing precise joint action:** Unlike retargeting methods, our exoskeleton reads precise joint angles directly from encoders, eliminating inaccuracies due to visual fingertip tracking.
- **Matching tactile information for learning:** Most handheld grippers for data collection [10, 42, 55] do not record the tactile information. Our design includes additional tactile sensors on the fingertip to record the same tactile info as what the robot hand would record.

Our **software adaptation** takes the form of a data processing pipeline that bridges the visual observation gap between human demonstration and robot deployment. This processing pipeline first removes the human hand and exoskeleton from the demonstration video using video segmentation, then

inpaints the video with the corresponding robot hand and environment backgrounds that match the target action. This adaptation ensures visual input consistency between training and robot deployment, despite visual differences between human and robotic hands.

With both hardware and software adaptation layers, DexUMI allows us to collect data on various tasks with minimal kinematic and visual gaps then transfer skills to robots. Comprehensive real-world experiments demonstrate DexUMI's capability on two different dexterous hand types: a 6-DoF Inspire hand [16] and a 12-DoF XHand [17]. Our approach achieves 3.2 times greater data collection efficiency compared to teleoperation and an average success rate of 86% across four tasks , including long-horizon and complex tasks requiring multi-finger contacts.

II. Related Work

Although extensive work has studied how to enable learning in simulated environments [3, 33, 61, 57, 39, 65, 22, 1, 24, 68, 30, 27, 45, 58, 38], we focus on reviewing real world data collection methods.

Teleoperation: Teleoperation is a popular interface for dexterous manipulation. Hand control is achieved with motion capture gloves [72, 56, 35, 53, 69], virtual-reality devices [26, 12, 9], or camera-based tracking [28, 66, 48, 21, 5, 23, 46]. Most approaches employ optimization-based retargeting to map human fingertips to robot hand. While being adaptable to different robot platforms, retargeting struggles with fundamental morphological differences between human and robot hands, especially the thumb flexibility [4]. Recent work by Zhou et al. [73] introduced a hand exoskeleton for direct joint mapping, but the mechanical structural differences limit the mapping accuracy. Additionally, teleoperation or kinesthetic teaching [25] require the robot hardware to be present, limiting the flexibility of data collection. In contrast, DexUMI collects manipulation data without physical robots.

Human hand video: Learning manipulation skills from human hand video is an attractive direction. Prior works have explored learning affordance [29, 40, 63, 15] or extracting human and object pose [43, 8, 41, 67, 47] from video. Though showing promising results, many of these works either require additional real-world robot data or need to learn the policy in simulation and depend on privileged information, such as object pose, to deploy the policy in the real world.

Wearable devices: Another line of work focuses on designing wearable devices for data collection, such as portable handheld grippers [59, 10, 20, 70, 13, 44, 52, 54, 42, 55, 32, 34, 37]. These approaches have demonstrated promising results in scaling real-robot manipulation skills. However, these systems primarily target simple parallel/pinch grippers and cannot be easily adapted to multi-fingered systems. Alternatively, Dexcap [60] uses motion capture gloves for in-contact data collection. However, it still relies on retargeting methods and human-correction data through teleportation. In contrast, our method eliminates these requirement, enabling direct policy deployment with data collected through DexUMI. Recently,

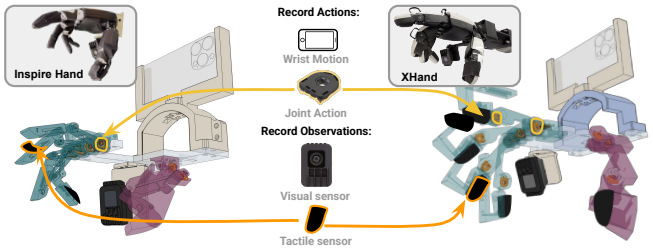

Fig. 2: **Exoskeleton Design.** The optimized exoskeleton design shares the same joint-to-fingertip position mapping as the target robot hand while maintaining the wearability. The exoskeletons utilizes the encoder to precisely capture the joint action and 150° DFoV camera to record the information-rich visual observation. An iPhone is rigidly mounted to track the wrist pose through the ARKit.

Wei and Xu [62] and Fang et al. [14] proposed hand-over-hand systems for dexterous hands. These works require the actual robot hand to be available and lifted by the human hand.

III. Hardware Adaptation to Bridge the Embodiment Gap

This section introduces our hardware adaptation, which is a wearable exoskeleton design that adapts human motion to feasible robot actions. While the final exoskeleton design is robot-specific, the principles of the design framework can be shared. We introduce the design framework in two parts: mechanism design optimization (§III-A) and sensor integration (§III-B).

A. Exoskeleton Mechanism Design

Modern robot hands often closely mimic human hands anatomically, meaning that a hand exoskeleton would compete for space with the human hand wearing it. The biggest challenge is for the thumb, whose pronation–supination movement can sweep a large volume and cause significant collision between the human thumb and a naively designed exoskeleton. Our exoskeleton design has two goals to achieve:

1) *Shared joint-action mapping:* The exoskeleton and the target robot hand must share the same joint-to-fingertip position mapping, including their limits, so the action can transfer.
2) *Wearability:* The exoskeleton must allow sufficient natural movements of the user's hand.

While the first goal can be mathematically defined, the wearability goal is hard to write down concretely. Our solution is to parameterize the exoskeleton design and formulate the wearability requirements as constraints on the design parameters, then find a solution that accommodates wearability while preserving kinematic relationships by solving an optimization. To make the optimization feasible, we prioritize the exact kinematics of fingertip links, while allowing greater flexibility in the kinematics of links less likely to contact objects.

E.1 Design initialization: We initialize the design with parameterized robot hand models based on URDF files (See Fig. 3). When such detailed designs are unavailable (e.g., the Inspire-Hand's finger mechanisms), we substitute them with equivalent general linkage designs with the same DoFs (e.g.,

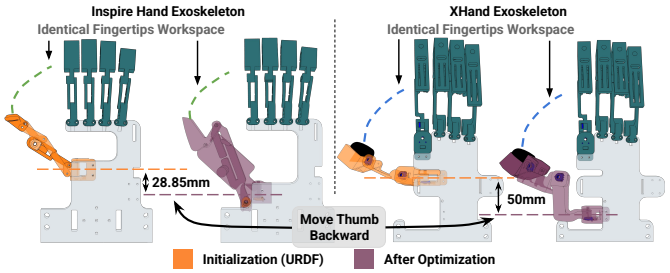

Fig. 3: **Mechanism Optimization.** To avoid thumb collision between human hand and exoskeleton, the hardware optimization step allows us to move the exoskeleton thumb backward while still preserving the original fingertip and joint mapping in SE(3) space.

a four-bar linkage) and allow optimization to find parameters that best match the observed kinematic behavior. Please see Appendix for details.

E.2 Bi-level optimization objective: Our optimization objective maximizes the following similarity: $\max_{\mathbf{p}} \mathcal{S}(\mathcal{W}^{\text{tip}}_{\text{exo}}(\mathbf{p}), \mathcal{W}^{\text{tip}}_{\text{robot}})$, where $\mathcal{W}^{\text{tip}}_{\text{exo}}$ and $\mathcal{W}^{\text{tip}}_{\text{robot}}$ represent the fingertip workspaces (set of all possible fingertip pose in SE(3)) for the exoskeleton and robot hand, respectively. $\mathbf{p} = \{j_1, ..., j_n, l_1, ..., l_m\}$ is the exoskeleton design parameters including joint positions $j_i \in \mathbb{R}^3$ in the wrist coordinate (i.e., flange) and linkage lengths l_j. The function $\mathcal{S}(\cdot, \cdot)$ represents a similarity metric between the two workspaces, which quantifies how closely the exoskeleton's fingertip pose distribution matches that of the robot hand. In practice, the $\mathcal{S}(\cdot, \cdot)$ is implemented as minimization by sampling configurations from both workspaces. Given a set of K robot hand configurations $\theta_{\text{robot},k}$ and N exoskeleton configurations $\theta_{\text{exo},n}$:

$$\begin{aligned}\mathcal{S}(\mathcal{W}^{\text{tip}}_{\text{exo}}(\mathbf{p}), \mathcal{W}^{\text{tip}}_{\text{robot}}) = \\ -\left(\sum_{k=1}^{K} \min_{\boldsymbol{\theta}_{\text{exo}}} \|\mathcal{F}^{\text{tip}}_{\text{exo}}(\mathbf{p}, \boldsymbol{\theta}_{\text{exo}}) - \mathcal{F}^{\text{tip}}_{\text{robot}}(\boldsymbol{\theta}_{\text{robot},k})\|^2 \right. \\ \left. + \sum_{n=1}^{N} \min_{\boldsymbol{\theta}_{\text{robot}}} \|\mathcal{F}^{\text{tip}}_{\text{exo}}(\mathbf{p}, \boldsymbol{\theta}_{\text{exo},n}) - \mathcal{F}^{\text{tip}}_{\text{robot}}(\boldsymbol{\theta}_{\text{robot}})\|^2 \right)\end{aligned} \tag{1}$$

where $\mathcal{F}^{\text{tip}}_{\text{exo}}$ and $\mathcal{F}^{\text{tip}}_{\text{robot}}$ are the forward kinematics for the exoskeleton and robot hand respectively. Optimizing the first term encourages the exoskeleton to cover the robot hand's workspace by finding exoskeleton configurations closest to the sampled robot hand configurations. The second term requires $\mathcal{W}^{\text{tip}}_{\text{exo}}(\mathbf{p}) \subseteq \mathcal{W}^{\text{tip}}_{\text{robot}}$, ensuring the exoskeleton's fingertip workspace remains within the robot hand's capabilities, preventing generation of unreachable poses outside the robot hand's workspace.

E.3 Constraints: We apply bound constraints $j_i \in \mathcal{C}_i$ and $l_j^{\min} \leq l_j \leq l_j^{\max}$, which are empirically selected to ensure that the exoskeleton can be comfortably worn. For example, we want to move the thumb swing joint closer to the wrist along the x-axis under MANO [51] convention to avoid collision between the human thumb's pronation–supination movement and that of the exoskeleton.

B. Sensor Integration

Sensors on the exoskeleton need to satisfy the following design objectives:

1) *Capture sufficient information:* the sensors need to capture ALL the information necessary for policy learning, which includes: robot action such as joint angle (*S.1*) and wrist motion (*S.2*), as well as observations in both vision (*S.3*) and tactile (*S.4*).
2) *Minimize embodiment gap:* the sensory information should have minimal distribution shift between human demonstration and robot deployment.

S.1 Joint capture & mapping. To precisely capture joint actions, our exoskeleton integrates joint encoders at every *actuated* joint – using resistive position encoders for both the XHand and Inspire-hand. We choose the `Alps` encoder [2] for its size and precision. Due to the joint friction and motor backlash, the mapping between exoskeleton joint encoder θ^i_{exo} and robot hand motor $\mathcal{M}^i_{\text{robot}}$ values is often non-linear, therefore, we train a simple regression model for each joint to obtain this mapping. To calibrate the regression model, we collect a set of paired data by uniformly sampling K motor values on the physical robot for each finger and then find the corresponding exoskeleton joint value by overlaying the visual observation between the robot hand and exoskeleton. This process creates a paired dataset for us to train the regression model.

S.2 Wrist pose tracking. We use iPhone ARKit to capture the 6DoF wrist pose, as smartphones represent the most accessible devices capable of providing precise spatial tracking. This tracking device is only needed for data collection, not for robot deployment.

S.3 Visual observation. We mounted a 150° diagonal field of view (DFoV) wide-angle camera `OAK-1` [19] under the wrist for both the exoskeleton and the target robot dexterous hand. This positioning was chosen to effectively capture hand-object interactions. Critically, the camera poses in the wrist frame were identical for the exoskeleton and the robot hand, which maintains visual consistency between training and deployment.

S.4 Tactile sensing. The wearable exoskeleton allows users to directly contact objects and receive haptic feedback. However, this human haptic feedback cannot be directly transferred to the robotic dexterous hand. Therefore, we install tactile sensors on the exoskeleton to capture and translate these tactile interactions. To ensure consistent sensor readings, we install the same type of tactile sensors on the exoskeleton as those used on the target robot hand. For XHand, we use the electro-magnetic tactile sensor that comes with the hand. For the Inspire-Hand, we install the same resistive tactile sensor `Force Sensitive Resistor` [18] for both the exoskeleton and the robot hand.

IV. Software Adaptation to Bridge the Visual Gap

Fig. 4 shows the visual gap between human demonstration (a) and robot deployment (h). To bridge this visual gap, we developed a data processing pipeline to adapt the demonstration image into what the robot will see as if the robot hand was collecting data. This adaptation uses off-the-shelf pretrained

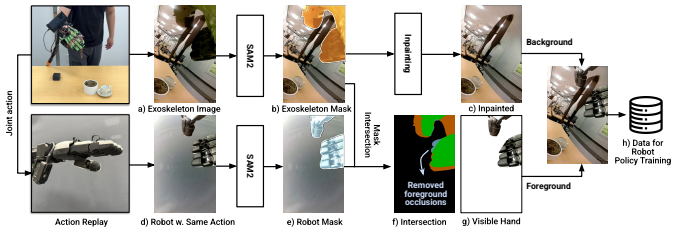

Fig. 4: **Bridging the Visual Gap.** To convert the visual observation into policy training data, we first segment the exoskeleton using SAM2 (b) and inpaint the missing background (c). The corresponding joint action (a) is replayed on the dexterous hand to obtain the robot hand image (d). SAM2 is applied to obtain the robot mask (e). The intersection (f) of the exoskeleton mask (b) and robot mask (e) identifies the visible part of the hand during interaction. Finally, we replace pixels in the inpainted background (c) with the visible robot hand (g).

models to ensure generalizability. The adaptation takes four steps:

V.1 Segment human hand and exoskeleton. Firstly, we segment (Fig. 4b) the human hand and exoskeleton on observation videos using SAM2 [49]. Since SAM2 requires initial prompt points, we established a protocol where the human operator always begins with the same hand gesture, allowing us to reuse the same prompt points for all demonstrations.

V.2 Inpaint environment background. With segmentation, we remove the human hand and the exoskeleton pixels from the image data. Then we use ProPainter [74], a flow-based inpainting method, to fully refill (Fig. 4c) the missing areas [6, 31, 7].

V.3 Record corresponding robot hand video. Next, to render robot hand properly into the video, we replay the recorded joint action on the robot hand and record another video with only the robot hand (Fig. 4d). This step does not involve the robot arm. We then used SAM2 again to extract the robot hand pixels (Fig. 4e) and discard the background. Notice, it is possible to train an image generation model to output the robot hand image based on the actions, but it requires additional model training.

V.4 Compose robot demonstrations. The last step is to merge the inpainted-background-only video with robot-hand-only video. It is crucial to maintain proper occlusion relationships: the robot hand does not always appear on top. We developed an occlusion-aware compositing approach leveraging: (1) our consistent under-wrist camera setup, and (2) the kinematic and shape similarity between the exoskeleton and robot hand. We compute a visible mask (Fig. 4f) by intersecting the exoskeleton mask and robot hand mask. Rather than naively overwriting pixels, we selectively replace pixels in the inpainted observation with robot hand pixels only if those pixels are present in the visible mask. This preserved natural occlusion relationships between the hand and objects when viewed from our under-wrist camera perspective. This approach generated visually coherent robot manipulation demonstrations that maintained proper spatial relationships.

Imitation learning. Our imitation learning policy $p(\mathbf{a_t}|o_t, f_t)$ takes processed visual observation o_t and tactile sensing f_t as input. The output is a sequence of actions $\{a_t, \ldots, a_{t+L}\}$ of length L, starting from the current time t, denoted as $\mathbf{a_t}$. The robot action a_t includes a 6-DOF end-effector action and N-DOF hand action where N depends on the specific robot hand hardware.

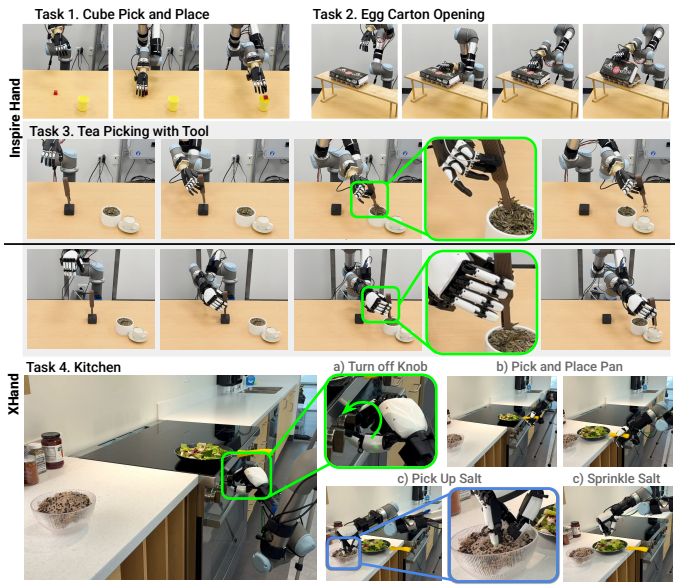

Fig. 5: **Policy Rollout:** We evaluate DexUMI's capabilities across challenging real-world tasks. The **Cube** task tests basic picking precision. The **Egg Carton** task evaluates multi-finger coordination. The **Tea Picking** task assesses performance on contact-rich manipulation requiring millimeter-level fine-grained fingertip actions. Finally, the **Kitchen** task tests capabilities on long-horizon high-precision actions to manipulate a knob, move a pan using both the side of thumb and index finger (beyond just fingertips), and utilize tactile sensing for visually challenging salt picking tasks.

V. Evaluation

Target robot hands: We evaluate DexUMI across two different robot hands:

- *Inspire Hand (IHand):* A twelve-DoF (six active DoFs) underactuated hand. The thumb has two active and two passive DoFs, while each remaining finger has one active and one passive DoF.
- *XHand:* A fully-actuated hand with twelve active DoFs. The thumb contains three DoFs, the index finger has three DoFs, and each of the remaining fingers has two DoFs.

Tasks: We evaluate DexUMI across four different real-world tasks:

- **Cube** [IHand]: Pick up a 2.5cm wide cube from a table and place it into a cup. This evaluates the basic capabilities and precision of the DexUMI system.
- **Egg Carton** [IHand]: Open an egg carton with multiple fingers: the hand needs the index, middle, ring, and little fingers to apply downward pressure on the carton's top while simultaneously using the thumb to lift the front latch.
- **Tea** [IHand & XHand]: Grasp tweezers from the table and use them to transfer tea leaves from a teapot to a cup. The main challenge is to stably operate the deformable tweezers with multi-finger contacts.

Method			Inspire Hand			
Action	Tactile	Visual	Cube	Carton	Tea	
					tool	leaf
Rel	Yes	Inpaint	**1.00**	0.85	**1.00**	0.85
Abs	Yes	Inpaint	0.10	0.35	0.80	0.00
Rel	No	Inpaint	0.95	**0.90**	**1.00**	**0.90**
Abs	No	Inpaint	0.90	0.85	0.90	0.60
Rel	No	Mask	0.60	0.10	0.90	0.50
Rel	No	Raw	0.20	0.05	0.85	0.05

TABLE I: **Evaluation Results (Inspire Hand).** Stage-wise accumulated success rates for different combinations of finger action representation (Absolute vs Relative), tactile feedback (Yes vs No), and visual rendering approaches (Inpaint vs Mask/Raw).

Method			XHand				
Action	Tactile	Visual	Tea		Kitchen		
			tool	leaf	knob	pan	salt
Rel	Yes	Inpaint	**1.00**	**0.85**	**0.95**	**0.95**	**0.75**
Abs	Yes	Inpaint	**1.00**	0.25	0.50	0.45	0.00
Rel	No	Inpaint	0.95	0.80	**0.95**	**0.95**	0.15
Abs	No	Inpaint	**1.00**	0.75	0.60	0.60	0.0
Rel	No	Mask	/	/	/	/	/
Rel	No	Raw	/	/	/	/	/

TABLE II: **Evaluation Results (XHand).** Stage-wise accumulated success rates. The Mask and Raw visual approaches were not tested on XHand (indicated by /).

- **Kitchen** [XHand]: The task involves four sequential steps: turn off the stove knob; transfer the pan from the stove top to the counter; pick up salt from a container; and lastly, sprinkle it over the food in the pan. The task tests DexUMI's capability over long-horizon tasks with precise actions, tactile sensing and skills beyond using fingertips.

Comparison: We evaluate the impact of policy action space choices, tactile sensing, and software adaptation on system performance.

- *Relative vs. Absolute finger action:* We compare the form of finger action trajectory: absolute position or relative trajectory proposed by [10]. We always use relative position for wrist action.
- *With vs. Without tactile sensing:* We trained policies with and without tactile sensor input.
- *With vs. Without software adaptation:* We examine two variants without software adaptation: (1) Mask, which replaces pixels occupied by the exoskeleton (during training) or robot hand (during inference) with a green color mask, and (2) Raw, which simply passes unmodified images containing the exoskeleton as policy input.

Evaluation protocol: For each evaluation episode, the test objects are randomly placed on the table at initialization. We conduct 20 evaluation episodes per task, maintaining consistent initial object configurations across our method and all baselines. For long horizon tasks, we report stage-wise accumulated success rate in Tab. I and Tab. II.

A. Key Findings

DexUMI framework enables efficient dexterous policy learning: As shown in Tab. I and Tab. II, the DexUMI system achieves high success rates across all four tasks on two robot hands. The system handles precise manipulation, long-horizon tasks, and coordinated multi-finger contact, while effectively generalizing across diverse manipulation scenarios.

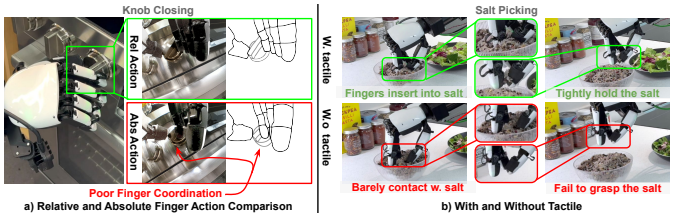

Fig. 6: **Comparisons.** a) The policy outputs relative hand actions yield more precise action and demonstrate better multi-finger coordination. Note, we draw a sketch for the knob closing for better visualization. b) Even with noisy tactile sensor reading, the tactile significantly improve tasks which is visually challenging.

Relative finger trajectories are more robust to noise and hardware imperfections: Tab. I and Tab. II show relative finger trajectory consistently achieves better success across all tasks. Fig. 6 shows more insights: relative trajectory can make critical contact events more reliable. We hypothesize two reasons for this difference: 1. Relative action has a simpler distribution than absolute and is thus easier to learn; 2. Relative action learns a reactive behavior where the delta action keeps accumulating until a key event is reached (e.g. fingers close on contact). However, the absolute action learns a static mapping and would stall if the mapping has errors.

Only relative finger trajectories can benefit from the noisy tactile feedback: An interesting observation in Tab. **??** is how having tactile affects the results differently. The tactile sensor on the XHand can drift and become inconsistent after experiencing high pressure. Therefore, in most cases, having tactile makes the results worse. We observed that only with relative trajectory can the policy benefit from having such tactile sensing. For the Inspire hand, the tactile sensors we manually installed are even more noisy (See section §III-B for details), then all methods become worse after adding tactile sensor as input. However, policies with relative trajectory still suffer less performance drop compared with the ones with absolute trajectory.

Tactile feedback improves performance on tasks with clean force profiles: We try to understand what kind of task would benefit from having tactile sensing. We focused on the XHand as its tactile sensors provide cleaner readings. We observed that tactile feedback significantly improved performance on picking up salt. This task highlights the effect of tactile because 1) The tactile sensors give a clear, large reading when the fingers touch the bowl of salt. 2) There is little useful visual information close to grasping as the camera view is mostly blocked by the bowl. In this case, we found that tactile feedback completely changes policy behavior. With tactile sensors, the fingers always insert into the salt first then close the fingers. Without tactile feedback, the fingers attempt to grasp the salt sometimes in the air. On the contrary, tactile info does not help in tweezer manipulation, which lacks strong correlation between hand motion and force feedback. Holding

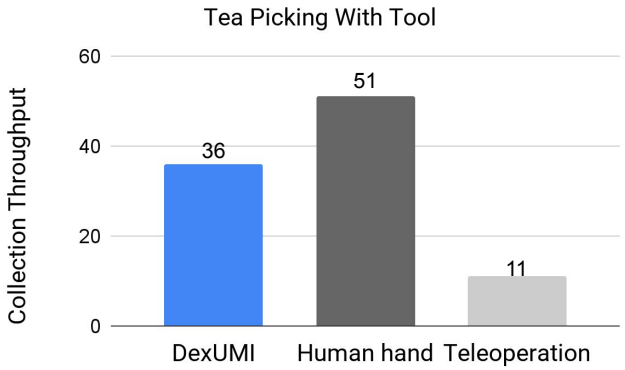

Fig. 7: **Efficiency:** Collection throughput (CT) within 15-minute. Though DexUMI still slower than bare hand, it achieves significant higher efficiency than teleportation.

a tweezer only triggers minimal tactile sensor readings.

DexUMI framework enables efficient dexterous hand data collection: We compared data collection efficiency across three ways: DexUMI, bare human hand, and teleoperation on the tea-picking-with-tool task. The same human operator collected data using each approach within 15-minute sessions. We computed the collection throughput (CT) based on the number of successful demonstrations acquired. As illustrated in Fig. 7, while DexUMI remains slower than direct human hand manipulation, it achieves 3.2 times greater efficiency than traditional teleoperation methods, significantly reducing the time required for dexterous manipulation data collection.

VI. Conclusion

We present DexUMI, a scalable and efficient data collection and policy learning framework that uses the human hand as an interface to transfer human hand motion to precise robot hand actions while providing natural haptic feedback. Through extensive challenging real-world experiments, we demonstrate DexUMI's capability in learning dexterous manipulation policies for precise, contact-rich, and long-horizon tasks. Our work establishes a new approach to collecting real-world dexterous hand data efficiently and at scale beyond traditional teleoperation.

VII. Limitation and Future Work

We would like to discuss DexUMI's limitations from three different aspects: hardware adaptation, software adaptation, and existing robot hand hardware.

Hardware Adaptation:

- *Per robot hand exoskeleton design:* Although DexUMI demonstrates generalizability across underactuated and fully-actuated hands, our optimization framework still requires hardware-specific tuning, especially for wearability. One future work direction is fully automated optimization formulation given robot hand model and some description of the human hand. Further, our hardware optimization framework can potentially leverage generative models [64] to increase efficiency and accuracy when design space grows.
- *Fingertips Matching:* Our current formulation focuses only on matching the fingertip workspace between the designed exoskeleton and target robot hand. It would be interesting for future work to also model remaining potential contact geometries such as the palm.
- *Wearability:* The hardware optimization pipeline makes the exoskeleton wearable and allows humans to operate it relatively easily for extended periods. However, wearability could be further improved by integrating soft materials, such as TPU for parts that contact the human hand. Additionally, constrained by both the design of the target hand and 3D printing material strength, users might still experience limitations in fully stretching certain fingers.
- *Reliability of Tactile Sensors:* Throughout our experiments, we found that reliable tactile sensors are key to maintaining consistent tactile observation between the exoskeleton and corresponding robot hand, thereby reducing the embodiment gap. In our implementation, the resistive tactile sensors added to the Inspire hand and its exoskeleton proved sensitive to their attachment way on fingers. Meanwhile, the electromagnetic tactile sensors on the XHand and its exoskeleton showed a tendency to drift after exposure to high pressure. Since the human hand generates more force than the robot hand, tactile sensor readings frequently drift when humans operate the exoskeleton. Future work can also incorporate other types of tactile sensors, such as vision-based tactile sensors [71, 50, 36] and capacitive F/T sensors [11].
- *Material Limitations:* Our experiments demonstrate that DexUMI is able to capture fine-grained fingertip actions such as closing tweezers. However, we sometimes found that encoders cannot precisely capture human motion due to 3D printing material strength limitations; occasionally, the human hand slightly distorts the exoskeleton linkage when manipulating objects. In such cases, encoders are unable to capture this distortion.

Software Adaptation:

- *Robot Hand Image:* Currently, we still require real-world robot hardware to obtain robot hand images. However, this requirement could be eliminated by implementing an image generation model that receives motor values as input and produces corresponding hand pose images as output.
- *Inpainting Quality:* Throughout our experiments, we found that the current software adaptation pipeline can already yield high-fidelity robot hand images. Nevertheless, we observed that illumination effects on the robot hand cannot be fully reproduced, and some areas in the image appear blurred due to limitations in the inpainting process.
- *Camera Location:* DexUMI currently requires the camera to be rigidly attached to the robot hand/exoskeleton and does not support a moving camera. However, it would be feasible to collect a dataset and train an image generation model that receives the relative pose between the camera and hand, along with hand pose information, to generate

the corresponding hand pose image from any given camera position.

Existing Robot Hand Hardware:

- *Precision:* Throughout our experiments, we found that both the Inspire Hand and XHand lack sufficient precision due to backlash and friction. For example, the fingertip location of the Inspire Hand differs when moving from 1000 to 500 motor units compared to moving from 0 to 500 motor units. Although the desired motor value is the same in both cases, the final fingertip position varies. We observed this phenomenon in both robot hands. Consequently, when fitting regression models between encoder and hand motor values, we can typically ensure precision in only "one direction"—either when closing the hand or opening it. This inevitably causes minor discrepancies in the inpainting and action mapping processes. Further, we found that the XHand mapping between motor command and fingertip location slightly differs across time shifts or after each reboot.
- *Size Discrepancy:* The size difference between the robot hand and the human hand may cause wearability issues. For example, if the robot hand is twice as large as the human hand, it becomes difficult for both the human hand and the exoskeleton to reach the joint configurations required by the robot hand.
- *Co-design:* Many of these wearability issues arise from design constraints in existing commercial hardware. An interesting direction would be to explore a reverse design paradigm: first designing an exoskeleton that is comfortable and fully operable for humans, and then using that exoskeleton as the foundation for designing the robot hand.

REFERENCES

[1] Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. arXiv preprint arXiv:2312.02975, 2023.

[2] Alps Alpine. Alps alpine rdc506018a rotary position sensor, 2025. URL https://www.digikey.com/en/products/detail/alps-alpine/RDC506018A/19529120. Accessed: March 23, 2025.

[3] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. The International Journal of Robotics Research, 39 (1):3–20, 2020.

[4] Sridhar Pandian Arunachalam, Irmak Güzey, Soumith Chintala, and Lerrel Pinto. Holo-dex: Teaching dexterity with immersive mixed reality. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 5962–5969, 2023. doi: 10.1109/ICRA48891.2023.10160547.

[5] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In 2023 ieee international conference on robotics and automation (icra), pages 5954–5961. IEEE, 2023.

[6] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. arXiv preprint arXiv:2207.09450, 2022.

[7] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. arXiv preprint arXiv:2409.03403, 2024.

[8] Zerui Chen, Shizhe Chen, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Vividex: Learning vision-based dexterous manipulation from human videos. arXiv preprint arXiv:2404.15709, 2024.

[9] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In 8th Annual Conference on Robot Learning.

[10] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. arXiv preprint arXiv:2402.10329, 2024.

[11] Hojung Choi, Jun En Low, Tae Myung Huh, Gabriela A Uribe, Seongheon Hong, Kenneth AW Hoffman, Julia Di, Tony G Chen, Andrew A Stanley, and Mark R Cutkosky. Coinft: A coin-sized, capacitive 6-axis force torque sensor for robotic applications. arXiv preprint arXiv:2503.19225, 2025.

[12] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. arXiv preprint arXiv:2407.03162, 2024.

[13] Kiran Doshi, Yijiang Huang, and Stelian Coros. On hand-held grippers and the morphological gap in human manipulation demonstration. arXiv preprint arXiv:2311.01832, 2023.

[14] Hao-Shu Fang, Branden Romero, Arthur Hu, Lirui Wang, Edward Adelson, and Pulkit Agrawal. Dexo: Hand exoskeleton system for teaching robot dexterous manipulation in-the-wild. 2023. URL https://fang-haoshu.github.io/files/DEXO.pdf.

[15] Alexey Gavryushin, Xi Wang, Robert JS Malate, Chenyu Yang, Xiangyi Jia, Shubh Goel, Davide Liconti, René Zurbrügg, Robert K Katzschmann, and Marc Pollefeys. Maple: Encoding dexterous robotic manipulation priors learned from egocentric videos. arXiv preprint arXiv:2504.06084, 2025.

[16] Generic. inspire hand, . URL https://inspire-robots.store/collections/the-dexterous-hands/products/the-dexterous-hands-rh56dfx-series?variant=42735794422004.

[17] Generic. Xhand, . URL https://www.robotera.com/en/goods1/4.html.

[18] Generic. Zd10-100 force sensitive resistor (fsr) pressure sensor, 2025. URL https://www.amazon.com/Pressure-ZD10-100-Resistance-Type-Resistor-Sensitive/dp/B07MHTWR1C/. Accessed: March 23, 2025.

[19] Generic. Oak-1 w ov9792, 2025. URL https://shop.luxonis.com/products/oak-1-w?variant=44051403604191. Accessed: March 23, 2025.

[20] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. arXiv preprint arXiv:2407.10353, 2024.

[21] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Transactions on Graphics (ToG), 39(4):87–1, 2020.

[22] Yunhai Han, Mandy Xie, Ye Zhao, and Harish Ravichandar. On the utility of koopman operator theory in learning dexterous manipulation skills. In Conference on Robot Learning, pages 106–126. PMLR, 2023.

[23] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 9164–9170, 2020. doi: 10.1109/ICRA40945.2020.9197124.

[24] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makoviichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, et al. Dextreme: Transfer of agile in-hand manipulation from simulation to reality. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 5977–5984. IEEE, 2023.

[25] Yifan Hou, Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Adaptive compliance policy: Learning approximate compliance for diffusion guided control. arXiv preprint arXiv:2410.09309, 2024.

[26] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. arXiv preprint arXiv:2412.14803, 2024.

[27] Binghao Huang, Yuanpei Chen, Tianyu Wang, Yuzhe Qin, Yaodong Yang, Nikolay Atanasov, and Xiaolong Wang. Dynamic handover: Throw and catch with bimanual hands. arXiv preprint arXiv:2309.05655, 2023.

[28] Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. Open teach: A versatile teleoperation system for robotic manipulation. arXiv preprint arXiv:2403.07870, 2024.

[29] Aditya Kannan, Kenneth Shaw, Shikhar Bahl, Pragna Mannam, and Deepak Pathak. Deft: Dexterous fine-tuning for real-world hand policies. arXiv preprint arXiv:2310.19797, 2023.

[30] Gagan Khandate, Siqi Shang, Eric T Chang, Tristan Luca Saidi, Yang Liu, Seth Matthew Dennis, Johnson Adams, and Matei Ciocarlie. Sampling-based exploration for reinforcement learning of dexterous manipulation. arXiv preprint arXiv:2303.03486, 2023.

[31] Puhao Li, Tengyu Liu, Yuyang Li, Muzhi Han, Haoran Geng, Shu Wang, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 573–580. IEEE, 2024.

[32] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. arXiv preprint arXiv:2410.18647, 2024.

[33] Toru Lin, Zhao-Heng Yin, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Twisting lids off with two hands. arXiv preprint arXiv:2403.02338, 2024.

[34] Fangchen Liu, Chuanyu Li, Yihua Qin, Ankit Shaw, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. arXiv preprint arXiv:2504.06156, 2025.

[35] Hangxin Liu, Zhenliang Zhang, Xu Xie, Yixin Zhu, Yue Liu, Yongtian Wang, and Song-Chun Zhu. High-fidelity grasping in virtual reality using a glove-based system. In 2019 International Conference on Robotics and Automation (ICRA), pages 5180–5186, 2019. doi: 10.1109/ICRA.2019.8794230.

[36] Yun Liu, Xiaomeng Xu, Weihang Chen, Haocheng Yuan, He Wang, Jing Xu, Rui Chen, and Li Yi. Enhancing generalizable 6d pose tracking of an in-hand object with tactile sensing. IEEE Robotics and Automation Letters, 9(2):1106–1113, 2023.

[37] Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. In 8th Annual Conference on Robot Learning, 2024.

[38] Tyler Ga Wei Lum, Albert H Li, Preston Culbertson, Krishnan Srinivasan, Aaron D Ames, Mac Schwager, and Jeannette Bohg. Get a grip: Multi-finger grasp evaluation at scale enables robust sim-to-real transfer. arXiv preprint arXiv:2410.23701, 2024.

[39] Tyler Ga Wei Lum, Martin Matak, Viktor Makoviychuk, Ankur Handa, Arthur Allshire, Tucker Hermans, Nathan D Ratliff, and Karl Van Wyk. Dextrah-g: Pixels-to-action dexterous arm-hand grasping with geometric fabrics. arXiv preprint arXiv:2407.02274, 2024.

[40] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In 2021 IEEE international conference on robotics and automation (ICRA), pages 6169–6176. IEEE, 2021.

[41] Priyanka Mandikal and Kristen Grauman. Dexvip:

Learning dexterous grasping with human hand pose priors from video. In Conference on Robot Learning, pages 651–661. PMLR, 2022.

[42] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. arXiv preprint arXiv:2112.01511, 2021.

[43] Sungjae Park, Seungho Lee, Mingi Choi, Jiye Lee, Jeonghwan Kim, Jisoo Kim, and Hanbyul Joo. Learning to transfer human hand skills for robot manipulations. arXiv preprint arXiv:2501.04169, 2025.

[44] Pragathi Praveena, Guru Subramani, Bilge Mutlu, and Michael Gleicher. Characterizing input methods for human-to-robot demonstrations. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 344–353. IEEE, 2019.

[45] Haozhi Qi, Ashish Kumar, Roberto Calandra, Yi Ma, and Jitendra Malik. In-hand object rotation via rapid motor adaptation. In Conference on Robot Learning, pages 1722–1732. PMLR, 2023.

[46] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. IEEE Robotics and Automation Letters, 7(4):10873–10881, 2022. doi: 10.1109/LRA.2022.3196104.

[47] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In European Conference on Computer Vision, pages 570–587. Springer, 2022.

[48] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In Robotics: Science and Systems, 2023.

[49] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.

[50] Branden Romero, Hao-Shu Fang, Pulkit Agrawal, and Edward Adelson. Eyesight hand: Design of a fully-actuated dexterous robot hand with integrated vision-based tactile sensors and compliant actuation. In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1853–1860. IEEE, 2024.

[51] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics, 36 (6):1–17, November 2017. ISSN 1557-7368. doi: 10.1145/3130800.3130883. URL http://dx.doi.org/10.1145/3130800.3130883.

[52] Felipe Sanches, Geng Gao, Nathan Elangovan, Ricardo V Godoy, Jayden Chapman, Ke Wang, Patrick Jarvis, and Minas Liarokapis. Scalable. intuitive human to robot skill transfer with wearable human machine interfaces: On complex, dexterous tasks. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 6318–6325. IEEE, 2023.

[53] Max Schwarz, Christian Lenz, Andre Rochow, Michael Schreiber, and Sven Behnke. Nimbro avatar: Interactive immersive telepresence with force-feedback telemanipulation. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5312–5319. IEEE, 2021.

[54] Mingyo Seo, H. Andy Park, Shenli Yuan, Yuke Zhu, and Luis Sentis. Legato: Cross-embodiment imitation using a grasping tool. IEEE Robotics and Automation Letters, 10 (3):2854–2861, March 2025. ISSN 2377-3774. doi: 10.1109/lra.2025.3535182. URL http://dx.doi.org/10.1109/LRA.2025.3535182.

[55] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. arXiv preprint arXiv:2311.16098, 2023.

[56] Kenneth Shaw, Yulong Li, Jiahui Yang, Mohan Kumar Srirama, Ray Liu, Haoyu Xiong, Russell Mendonca, and Deepak Pathak. Bimanual dexterity for complex tasks. arXiv preprint arXiv:2411.13677, 2024.

[57] Leon Sievers, Johannes Pitz, and Berthold Bäuml. Learning purely tactile in-hand manipulation with a torque-controlled hand. In 2022 International conference on robotics and automation (ICRA), pages 2745–2751. IEEE, 2022.

[58] Ritvik Singh, Arthur Allshire, Ankur Handa, Nathan Ratliff, and Karl Van Wyk. Dextrah-rgb: Visuomotor policies to grasp anything with dexterous hands. arXiv preprint arXiv:2412.01791, 2024.

[59] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. Robotics and Automation Letters, 2020.

[60] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation, 2024. URL https://arxiv.org/abs/2403.07788.

[61] Jun Wang, Ying Yuan, Haichuan Che, Haozhi Qi, Yi Ma, Jitendra Malik, and Xiaolong Wang. Lessons from learning to spin" pens". arXiv preprint arXiv:2407.18902, 2024.

[62] Dehao Wei and Huazhe Xu. A wearable robotic hand for hand-over-hand imitation learning. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 18113–18119. IEEE, 2024.

[63] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In Conference on Robot Learning, pages 618–629. PMLR, 2023.

[64] Xiaomeng Xu, Huy Ha, and Shuran Song. Dynamics-guided diffusion model for robot manipulator design. arXiv preprint arXiv:2402.15038, 2024.

[65] Max Yang, Chenghua Lu, Alex Church, Yijiong Lin, Chris Ford, Haoran Li, Efi Psomopoulou, David AW Barton, and Nathan F Lepora. Anyrotate: Gravity-invariant in-hand object rotation with sim-to-real touch. arXiv preprint arXiv:2405.07391, 2024.

[66] Shiqi Yang, Minghuan Liu, Yuzhe Qin, Runyu Ding, Jialong Li, Xuxin Cheng, Ruihan Yang, Sha Yi, and Xiaolong Wang. Ace: A cross-platform and visual-exoskeletons system for low-cost dexterous teleoperation. In 8th Annual Conference on Robot Learning.

[67] Jianglong Ye, Jiashun Wang, Binghao Huang, Yuzhe Qin, and Xiaolong Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. IEEE Robotics and Automation Letters, 8(5):2882–2889, 2023.

[68] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. arXiv preprint arXiv:2303.10880, 2023.

[69] Zhao-Heng Yin, Changhao Wang, Luis Pineda, Francois Hogan, Krishna Bodduluri, Akash Sharma, Patrick Lancaster, Ishita Prasad, Mrinal Kalakrishnan, Jitendra Malik, et al. Dexteritygen: Foundation controller for unprecedented dexterity. arXiv preprint arXiv:2502.04307, 2025.

[70] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. In Conference on Robot learning, pages 1992–2005. PMLR, 2021.

[71] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors, 17(12):2762, 2017.

[72] Han Zhang, Songbo Hu, Zhecheng Yuan, and Huazhe Xu. Doglove: Dexterous manipulation with a low-cost open-source haptic force feedback glove, 2025. URL https://arxiv.org/abs/2502.07730.

[73] Jianshu Zhou, Boyuan Liang, Junda Huang, Ian Zhang, Pieter Abbeel, and Masayoshi Tomizuka. Global-local interface for on-demand teleoperation. arXiv preprint arXiv:2502.09960, 2025.

[74] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10477–10486, 2023.