

ViHaTeleop: A Low-Cost, Lightweight Visual-Haptic Teleoperation System for Dexterous Manipulation Learning

Abstract—Learning from demonstration is a promising approach for dexterous manipulation, but collecting high-quality contact-critical demonstrations remains difficult with low-cost teleoperation hardware. We present ViHaTeleop, a lightweight (0.7 kg), low-cost (\$550) visual-haptic teleoperation system with SLAM-based wrist tracking, camera-based hand tracking, and finger-wise vibrotactile feedback through Linear Resonant Actuators (LRA). The system includes several design choices (LED illumination, fisheye hand camera, and tactile-aware retargeting constraints) and is deployed on Franka + LEAP Hand + 9DTact in both real and simulated environments. Under matched with/without-haptic conditions with nine participants across six contact-critical tasks, haptics improved success rates across all tasks (+2.2 to +15.6 percentage points), while completion-time effects were task-dependent. Subjective ratings showed significant gains in contact clarity and grasp confidence in both simulation and real-world settings (Wilcoxon signed-rank, $p < 0.05$). We also integrate a lightweight depth-camera-based tactile proxy in Isaac Sim, enabling a full pipeline from multi-modal demonstration collection to visual-tactile policy training. Preliminary downstream validation by training visual-tactile policies from collected demonstrations shows tactile cues benefit contact-critical subtasks (peg-in-hole: +17 percentage points over vision-only).

I. INTRODUCTION

Large-scale robot manipulation data is critical for embodied AI, with teleoperation-based data collection playing an increasingly important role [1], [2], [3], [4], [5]. However, current teleoperation systems face a fundamental limitation: expensive systems (>\$10k) provide haptic feedback for contact-critical manipulation, while affordable systems (<\$1k) lack haptic feedback entirely. This gap particularly hinders data collection for fine-grained manipulation tasks requiring delicate force control—such as grasping fragile objects (e.g., potato chips) or performing precise assembly—where haptic feedback is essential for operator awareness and performance.

In this paper, we define contact-critical tasks as tasks where (i) success depends on detecting contact onset and maintaining stable contact, and (ii) errors are strongly coupled to interaction forces (e.g., slip, breakage, jamming, or insertion failure). Visual constraints are treated as an evaluation context rather than part of the definition.

Our primary research question is whether low-cost vibrotactile feedback measurably improves contact-critical teleoperation quality in a matched vision-based setup. While several systems incorporate haptic feedback [6], [7], the impact of low-cost vibrotactile actuators has not been systematically quantified under matched with/without-haptic conditions. In our current 9-participant study, haptic feedback improved success rates across all tasks, with larger gains on several

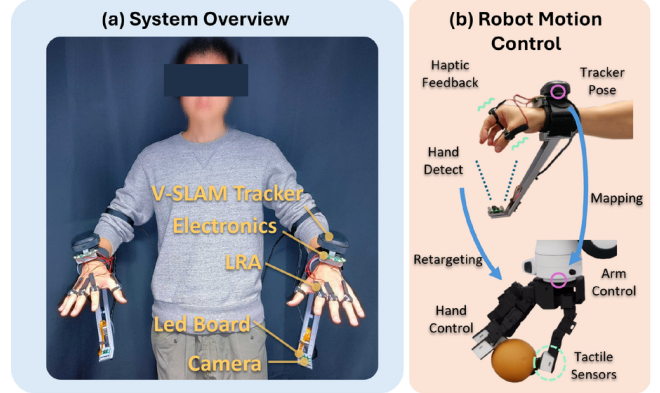


Fig. 1. Hardware overview and control architecture of ViHaTeleop. (a) The system consists of two bimanual wrist and hand configuration tracking modules. Each module uses a VSLAM+IMU tracker (HTC Vive Ultimate) for wrist pose tracking and a fisheye camera for hand configuration detection in real-time. LRA ring on each finger provides vibrational haptic feedback. (b) The captured wrist and hand configurations are mapped and retargeted to control a robotic arm (Franka) with a dexterous hand (LEAP Hand). Tactile sensors (9DTact) on the robot fingertips provide touch sensory feedback to the operator.

delicate tasks, while completion-time effects were mixed across tasks. As a secondary objective, we evaluate whether demonstrations collected with this system are usable for downstream visual-tactile imitation learning.

ViHaTeleop (illustrated in Fig. 1) consists of an off-the-shelf VSLAM+IMU-based tracker (HTC Vive Ultimate [8]) for capturing wrist motion, a fisheye camera (160-degree field of view, \$11) for detecting hand configuration, and LRAs (\$1.4 each) for generating vibrational haptic feedback. The camera is wrist-mounted, moving together with the hand—similar to ACE [9]—but with significantly reduced mass and torque. During operation, we capture the tracker pose and hand configuration in real-time. The tracker pose is mapped to the robot arm end-effector pose to control the Franka [10] arm via inverse kinematics with joint position commands. The hand configuration is retargeted to the LEAP Hand[11] joints using the method from [12]. The 9DTact [13] sensors (a low-cost, open-source tactile sensor) mounted on the LEAP Hand fingertips capture real-time tactile signals, which are converted to PWM control signals for the LRAs via a driver board mounted on the operator’s wrist. We implement ViHaTeleop to teleoperate a Franka + LEAP Hand + 9DTact platform both in reality and simulator. In simulation, we additionally provide a lightweight 9DTact approximation in Isaac Sim 5.0 [14] using a depth-camera-based proxy.

As shown in Table I, ViHaTeleop achieves a favorable balance across cost, weight, haptic feedback, and occlusion robustness compared to recently proposed systems. Through

TABLE I
COMPARISON OF TELEOPERATION SYSTEMS

System	Cost (USD)	Wearable Burden (kg)	Haptic Feedback	Occlusion Robust
NuExo [6]	>10k	Heavy (5.2)	✓	✓
BunnyVP [7]	4k	Light (0.65)*	✓	✗
DexCap [15]	3.6k	Light (1.8)†	✗	✓
AnyTeleop [12]	50	None	✗	✗
ACE [9]	600	Medium	✗	✓
Ours	550	Light (0.7)	✓	✓

*Vision Pro headset weight. †Full backpack system; glove portion is lighter.

targeted design improvements (LED lighting for varied illumination, a fisheye camera to reduce mounting lever arm and torque, intuitive tactile feedback, and retargeting constraints for flat-surface tactile sensors), ViHaTeleop provides a practical operating point for contact-critical data collection.

While ViHaTeleop builds upon established teleoperation techniques, we summarize our key contributions as follows:

- 1) Primary: controlled evaluation of low-cost vibrotactile feedback under matched with/without-haptic conditions across contact-critical tasks (user study).
- 2) A complete haptic integration pipeline: tactile-aware retargeting constraints (enabling multi-modal data quality), multi-fingertip 9DTact sensing, and LRA vibrotactile feedback mapping, on a low-cost (\$550), lightweight (0.7 kg) platform.
- 3) A simulation-integrated pipeline covering the full loop from multi-modal data collection to policy evaluation: Isaac Sim task environments, a lightweight depth-camera-based 9DTact proxy, and multi-fingertip tactile point cloud registration for the LEAP Hand. Preliminary validation using an extended 3D-ViTac architecture (first applied to >20-DoF dexterous manipulation) demonstrates usable visual-tactile demonstrations for downstream imitation learning.

The scope of this paper is primarily the teleoperation system study; the imitation-learning section is included as a constrained downstream feasibility validation rather than a standalone algorithmic contribution.

II. RELATED WORK

A. Dexterous Teleoperation

Hand Teleoperation: Dexterous hand teleoperation encompasses diverse approaches, each presenting distinct trade-offs. Vision-based methods using a single camera only [12], [16] or VR headsets [17], [7] are low-cost and lightweight but struggle with occlusion issues. Optical tracking systems [18], while more accurate, remain costly, non-portable, and still susceptible to occlusion despite multi-camera setups. Wearable sensor approaches [15], Manus®, Dexmo®, SenseGlove®—circumvent occlusion entirely by directly measuring joint angles. However, they are expensive and difficult to add customization for other feedback. Custom exoskeletons [19], [20], [21] physically replicate robot kinematics, enabling intuitive control at the expense of substantial design effort, limited cross-platform generalizability, and considerable operator burden—particularly when human and robot morphologies differ. With the anti-occlusion improvement

from [9], and hand detecting algorithms [22], [23] and efficient retargeting algorithm [16], [12], [7] mapping human hand configuration to various anthropomorphic robot hands (LEAP Hand[11], etc), the vision-based method works as a promising method for teleoperating dexterous robot hands to meet mass data collection purposes. DexPilot [16] introduced distance-dependent projection mechanisms for precise fingertip control, which we adopt as our retargeting foundation. We extend this framework with tactile-aware constraints tailored for flat-surface sensors (Sec. III-D).

Wrist Teleoperation: Fixed kinematic linkages [24], [25] provide stable and accurate tracking but lack portability. Wearable exoskeletons [26], [27], [9] improve portability but introduce considerably greater weight compared to tracker-based alternatives. Recent evaluation of SLAM-based tracker HTC Vive Ultimate [28] demonstrates performance competitive with or superior to low-cost kinematic systems using 3D-printed components and encoders, while maintaining minimal operator burden and high portability.

Teleoperation with Haptic Feedback: Integrating haptic feedback to improve teleoperation and the sense of tele-existence is a long-standing research area. Various mechanisms have been explored. Vision-only proxies for touch [29], [30] can help collect delicate manipulation data; however, extended reliance on visual perception alone is cognitively demanding and less intuitive for operators. Motorized force-feedback devices [31], [32], [19]—including off-the-shelf options such as Dexmo® and SenseGlove®—provide richer cues but are too expensive and too heavy. Electrotactile methods [33], [34] offer high spatial resolution, yet their fabrication processes are not mature enough for mass production. By contrast, low-cost, robust vibrotactile feedback [7] can improve performance without increasing operator burden. Moreover, a proper ring-mounted design is compatible with vision-based methods. We further conduct a user study to validate the haptic feedback integrated into our system, using thorough objective and subjective metrics.

While systems like NuExo [6] and ACE [9] provide valuable capabilities, each presents distinct limitations for large-scale data collection: NuExo’s weight (5.2 kg) and cost (>\$10k) hinder prolonged use and mass deployment, while ACE’s exoskeleton adds considerable weight, and its mobile version suffers from unstable end-effector mapping due to a non-tracked floating base. We adopt ACE’s hand-tracking approach but replace the exoskeleton with SLAM-based tracking, providing a fixed-origin reference frame for more intuitive and stable mapping. Importantly, ViHaTeleop maintains interface consistency with ACE [9], enabling potential extension to other robotic platforms beyond Franka + LEAP Hand.

B. Visual-Tactile Policy Learning from Human Demonstrations

Though still under-explored, visual-tactile policy learning from human demonstrations is gaining increasing attention, with several recent works integrating tactile modality alongside vision [29], [30], [19], [18]. When combined with

high-DoF dexterous hands [11] and high-resolution tactile sensors [35], [13], policy learning becomes a challenging high-dimensional problem. The imitation learning method 3D-ViTac [30] offers a promising approach by integrating visual and dense-array tactile data into a unified 3D point cloud representation, demonstrating strong potential for dexterous hand manipulation. Given the high-quality, contact-critical, long-horizon manipulation data collected by our teleoperation system, we extend 3D-ViTac as a downstream learning module to evaluate our complete pipeline from data collection to policy execution.

III. SYSTEM DESIGN

The primary design targets are low cost (<\$550 using off-the-shelf components), lightweight wearable (0.7 kg via hollow 3D-printed linkages and fisheye camera), robustness (LED lighting for varied illumination), portability (modular snap-fit assembly in minutes), and haptic feedback integration—all essential for scalable contact-critical data collection.

A. Hardware Design

The hardware consists of four primary modules (Fig. 2): (1) *Tracker module*: HTC Vive Ultimate tracker for 6-DoF wrist pose tracking via VSLAM and IMU, communicating wirelessly over Wi-Fi; (2) *Camera module*: OV5647 with 160° fisheye lens for hand configuration detection, connected via USB; (3) *LED module*: Ring-mounted LEDs powered through USB for illumination enhancement; (4) *Haptic module*: LRA rings on each finger driven by a custom PCB with ESP32 microcontroller, connected via USB. We designed a custom detachable connector for the bottom-left interface of the Vive Tracker, enabling quick, tool-free attachment to the 3D-printed mount. The connector ensures stable fixation and accurate pose alignment while allowing easy removal for charging or reconfiguration. Combined with magnetic snap-fit joints between printed parts, all ViHaTeleop submodules can be detached and reassembled within minutes, enhancing portability and user convenience.

B. Pose Estimation

The operator’s wrist pose is captured by the tracker via VSLAM and IMU sensing, providing 6-DoF tracking. Hand images from the wrist-mounted fisheye camera are undistorted using pre-calibrated intrinsic parameters and processed with MediaPipe [23] to detect 21 hand keypoints in real-time. Integrated LED lighting activates under suboptimal illumination conditions. Figure 4 shows qualitative robustness improvements in challenging scenarios (direct light toward camera and low-light environments), helping stabilize pose estimation without controlled lighting. A dedicated quantitative tracker/camera accuracy benchmark is left for future work.

C. Robot Arm Control

The tracker provides both position and orientation. We map the tracker motion to a robot end-effector target pose

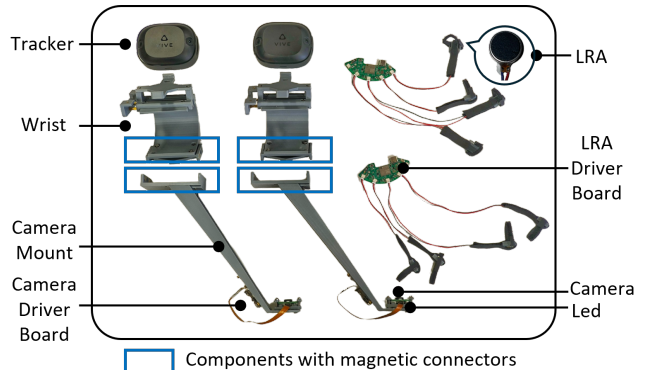


Fig. 2. Hardware Components.

using relative pose mapping: the operator’s initial wrist pose defines a reference frame, and subsequent wrist motion is computed as a relative transform. To handle workspace mismatch, we clamp target translation to a predefined workspace box in the robot base frame. The target pose is executed through inverse kinematics with joint limits; if IK fails or violates limits, we hold the last valid command to avoid discontinuities. The initial robot configuration is set to a fixed nominal pose, and the initial tracker pose is used as the teleoperation reference.

D. Robot Hand Motion Retargeting

DexPilot retargets human keypoints to LEAP Hand joint angles via optimization, encouraging fingertip geometry matching while regularizing joint motion. A distance-dependent projection/weighting increases fingertip emphasis as the thumb approaches a finger to improve pinch precision. We smooth this schedule because abrupt thresholding can cause discontinuous intermediate grasp configurations (“snapping”).

Tactile-Aware Parallelism Constraint: To maximize contact quality with flat-surface tactile sensors (e.g., 9DTact [13], GelSight [35]), we extend DexPilot’s framework by introducing a fingertip parallelism constraint. Standard retargeting optimizes for position matching but does not enforce orientation alignment, often resulting in edge contact that limits tactile sensing effectiveness. We augment the cost function with an alignment term:

$$\mathcal{L}_{\text{parallel}} = w_p \cdot \sum_{j \in \mathcal{F}} \|n_{\text{thumb}} \cdot n_j + 1\|^2 \quad (1)$$

where $\mathcal{F} = \{\text{index, middle, ring}\}$, and n_j denotes the outward surface normal of fingertip j computed from auxiliary URDF links. To prevent abrupt "snapping" in standard DexPilot projection, we extend its projection mechanism with adaptive functions that smoothly interpolate based on thumb-to-finger distance d :

$$\begin{aligned}\eta(d) &= \eta_{\min} + \frac{d - d_0}{d_{\max} - d_0}(\eta_{\max} - \eta_{\min}) \\ w(d) &= w_{\min} + \frac{d_{\max} - d}{d_{\max} - d_{\min}}(w_{\max} - w_{\min})\end{aligned}\quad (2)$$

where $\eta(d)$ controls projection distance and $w(d)$ modulates retargeting strength. For the LEAP Hand, we set $d_{\min} =$

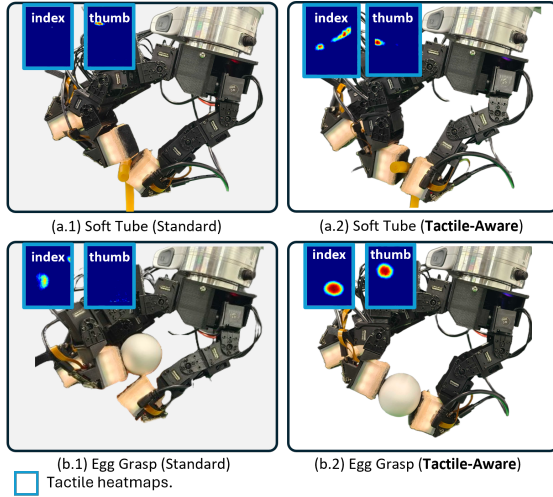


Fig. 3. Comparison of standard DexPilot retargeting vs. our tactile-aware extension on two manipulation tasks. (a) Soft tube grasping, (b) Egg grasping. Standard retargeting (left column) produces edge contact with negligible sensor signals, while our constraint (right column) enforces parallel fingertip alignment, enabling effective tactile sensing. Blue insets show tactile sensor heatmaps from thumb and index fingertips.

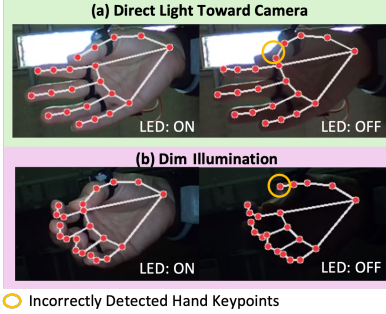


Fig. 4. Hand skeleton detection in challenging lighting: (a) direct glare toward camera; (b) low-light scene. LED illumination stabilizes detection. $d_0 = 30\text{mm}$, $d_{\max} = 50\text{mm}$, $\eta_{\min} = 0.1\text{mm}$, $\eta_{\max} = d_{\max}$, $w_{\min} = 1.0$, $w_{\max} = w_p = 200$. The parallelism weight w_p activates when $d < d_{\max}$, enabling smooth fingertip approach with intermediate grasp configurations.

The parallelism constraint can be toggled on/off before teleoperation based on task requirements—enabled for contact-critical tasks (pinching, rolling) and disabled for tasks requiring varied finger configurations (power grasps, hook grasps). As shown in Fig. 3, standard position-only retargeting produces edge contact that yields negligible tactile signals, while our tactile-aware constraint enforces parallel alignment that enables effective tactile sensing. Tactile sensor heatmaps (blue insets) confirm this improvement: standard retargeting shows minimal sensor activation, whereas our extension achieves robust contact engagement across both soft tube and egg manipulation tasks.

E. Vibrotactile Haptic Feedback

The haptic feedback system converts tactile signals into vibrotactile cues through the following processing pipeline. For each fingertip i , let $\delta_i^{\max}$ denote the maximum deformation depth measured by the 9DTact sensor array. The PWM command u_i for the corresponding LRA is computed as:

$$u_i = \text{LPF} \left(\text{PWM}_{\max} \cdot \frac{\delta_i^{\max} - \delta_{\min}}{\delta_{\max} - \delta_{\min}} \right) \quad (3)$$

where $[\delta_{\min}, \delta_{\max}]$ are pre-calibrated deformation thresholds, and LPF denotes a first-order low-pass filter for signal smoothing. This mapping is monotonic in deformation depth and provides an intuitive contact-intensity cue; proportionality to force holds only approximately under an elastic-contact assumption.

We mount LRAs on ring fixtures on the finger (rather than on the fingertip) to avoid interfering with fingertip contacts and vision-based tracking, while maintaining consistent placement and comfort. Despite the offset from the fingertip, users can localize finger-wise vibration due to spatial separation of stimulation sites.

We drive each LRA with a fixed carrier frequency set to the actuator’s resonance (from its datasheet) and modulate perceived intensity via PWM duty cycle. The parameters ($\text{PWM}_{\max}$ and depth thresholds) are chosen via short pilot trials to ensure perceptible but comfortable feedback and to avoid saturation during typical contacts.

IV. EXPERIMENTS

A. Evaluation Overview

We evaluate ViHaTeleop through two complementary studies: (1) a controlled ablation comparing teleoperation performance with and without haptic feedback, and (2) imitation learning experiments assessing whether collected demonstrations support downstream policy training. Our experiments focus on contact-critical manipulation tasks.

B. Implementation

Real Robot Setup. Franka FR3 (7-DoF) equipped with LEAP Hand (16-DoF) and 9DTact tactile sensors on three fingertips (thumb, index and middle fingers). An Intel RealSense D455 RGB-D camera is mounted in front of the workspace. Control runs at 20 Hz.

Simulation Setup. Isaac Sim 5.0 [14] with both single-arm and dual-arm Franka FR3 configurations. A depth-based proxy approximates 9DTact tactile feedback for efficient simulation.

C. User Study: Haptic Feedback Ablation

Participants. Nine volunteers (A–I) participated in the study. The sample included six males and three females, aged 22–31 years (mean = 26.3, SD = 3.39). Eight participants were right-handed. Regarding prior teleoperation experience, two reported extensive experience, one reported limited experience, and six reported no prior experience. Demographics are summarized in Table II.

We determined our sample size based on effect sizes observed in comparable haptic teleoperation studies. A priori power analysis (G*Power 3.1, Wilcoxon signed-rank test, two-tailed, $\alpha = 0.05$, power = 0.80) indicated that nine participants provide adequate power to detect large effects (Cohen’s $d \geq 0.8$), which are typical in haptic feedback ablation studies where differences are perceptually salient.

Tasks. We selected tasks to cover common contact-critical subskills: fragile grasping (egg, chip), insertion/contact on-set (peg-in-hole), multi-contact manipulation (stacking), and

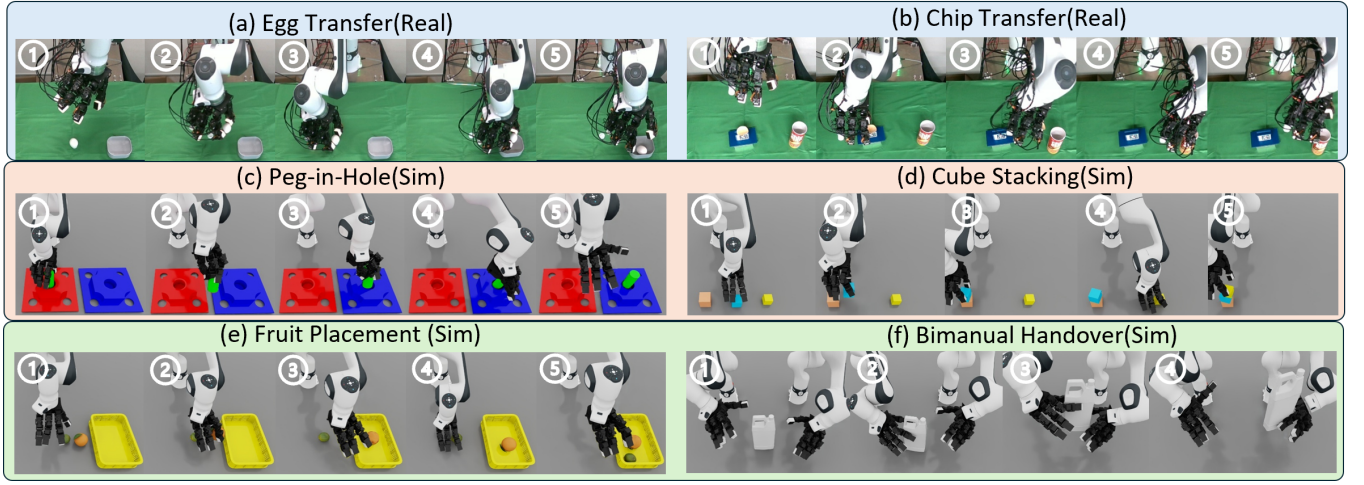


Fig. 5. Experimental tasks for user study. Numbers indicate key stages in task progression. Real-world: (a) egg transfer to bowl, (b) fragile chip grasping. Simulation: (c) peg-in-hole insertion, (d) three-cube stacking, (e) fruit pick-and-place, (f) bimanual bottle handover.

TABLE II
PARTICIPANT DEMOGRAPHICS (N=9)

Attribute	Summary
Age	22–31 (mean=26.3, SD=3.39)
Gender	6M (A–E, I), 3F (F–H)
Handedness	8R, 1L (C)
Experience	Experienced: B, D; Some: A; None: rest

bimanual coordination (handover). Real-robot tasks are performed co-located with direct line-of-sight, while simulation uses a fixed monitor view to emulate camera-based demonstration collection in embodied-AI data factories. Six contact-critical manipulation tasks requiring precise force modulation (Fig. 5):

Real World (operator standing beside robot):

- Egg transfer: Pick up an egg and place it into a bowl
- Chip transfer: Pick up a fragile potato chip and place it into a tube container

Simulation (operator viewing Isaac Sim GUI on monitor):

- Peg-in-hole: Extract a tube from one peg hole and insert it into another peg hole
- Cube stacking: Stack three cubes sequentially to form a tower
- Fruit pick-and-place: Pick an orange and place in basket, then pick a lime and place in the same basket
- Bimanual handover: Transfer a liquid container from right hand to left hand

Protocol. After a 10-minute tutorial and practice session with ViHaTeleop on both real and simulated platforms, participants practiced each task until achieving at least one successful trial to avoid floor effects. They then performed 10 experimental trials per task: 5 without haptic feedback and 5 with haptic feedback, presented in randomized order to control for learning effects.

Metrics.

- *Objective:* (1) Success rate per condition (successful trials out of 5); (2) Completion time for successful trials.
- *Subjective:* Post-condition 5-point Likert scale ratings

TABLE III
USER STUDY: SUCCESS RATES

Task	w/o Haptic	w/ Haptic	Δ	p
<i>Real World</i>				
Egg transfer	66.7%	80.0%	+13.3%	.172
Chip grasping	46.7%	48.9%	+2.2%	.516
<i>Simulation</i>				
Peg-in-hole	42.2%	51.1%	+8.9%	.250
Cube stacking	24.4%	28.9%	+4.5%	.500
Fruit placement	46.7%	57.8%	+11.1%	.531
Handover	51.1%	66.7%	+15.6%	.125

TABLE IV
USER STUDY: COMPLETION TIMES (SECONDS)

Task	w/o Haptic	w/ Haptic	Δ	p
<i>Real World</i>				
Egg transfer	27.1	23.6	−12.8%	.742
Chip grasping	26.9	28.5	+5.9%	1.00
<i>Simulation</i>				
Peg-in-hole	35.5	31.3	−12.0%	.078
Cube stacking	55.0	62.6	+13.8%	.359
Fruit placement	37.9	39.6	+4.4%	.887
Handover	21.6	30.4	+40.2%	.297

(1=strongly disagree, 5=strongly agree) for three dimensions:

- *Ease of manipulation:* “It was easy to manipulate objects using the teleoperation system.”
- *Contact clarity:* “I could clearly tell when the robot hand made contact with an object.”
- *Grasp confidence:* “I felt confident that the object was securely grasped in the robot hand.”

Statistics: Paired comparisons use the two-tailed Wilcoxon signed-rank test (significance at $p < 0.05$). Per-task statistical power is limited at $n = 9$; we report p -values for transparency rather than claiming per-task significance.

D. User Study Results

Success Rates. Table III shows descriptive success-rate trends across all tasks. In this 9-participant dataset, haptic feedback increased success rates on all six tasks, with

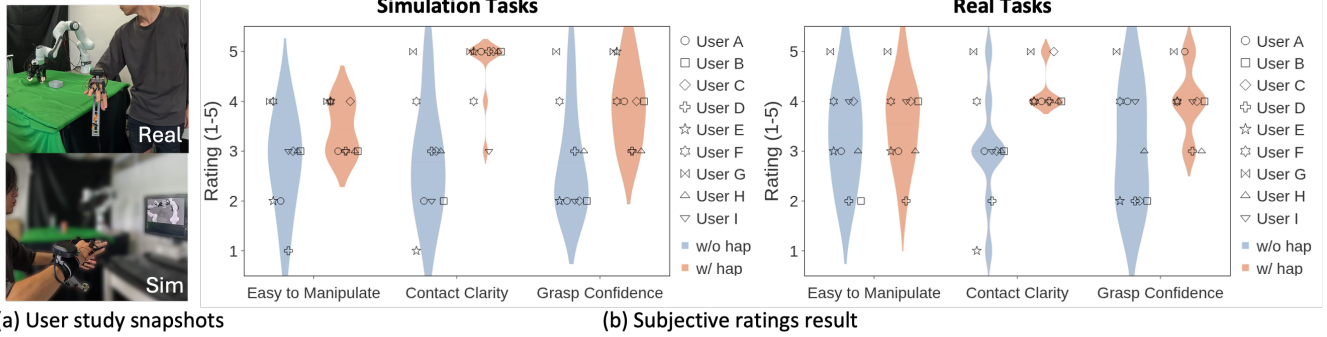


Fig. 6. User study experimental setup and subjective ratings. (a) Experimental snapshots showing teleoperation in real-world (top) and simulation (bottom) environments. (b) Subjective ratings comparing without and with haptic feedback across nine participants. Error bars indicate standard deviation. Statistical comparisons use two-tailed Wilcoxon signed-rank tests; detailed values are reported in Sec. IV-D.

TABLE V
SIMULATION CUBE SUCCESS RATE ABLATION: DEFAULT VS.
PARALLELISM CONSTRAINT

Haptic	Cube	Default	Parallelism	Δ
w/o	Blue	35.0%	50.0%	+15.0%
	Yellow	5.0%	10.0%	+5.0%
w/	Blue	55.0%	75.0%	+20.0%
	Yellow	5.0%	5.0%	0.0%

gains ranging from +2.2 to +15.6 percentage points. The largest gains were observed in simulation handover (+15.6 points) and real-world egg transfer (+13.3 points), while chip grasping showed a small gain (+2.2 points).

Completion Times. Table IV shows task-dependent effects on completion time. Completion time decreased for egg transfer (27.1s $\rightarrow$ 23.6s, -12.8%) and peg-in-hole (35.5s $\rightarrow$ 31.3s, -12.0%), but increased for chip grasping (+5.9%), cube stacking (+13.8%), fruit placement (+4.4%), and handover (+40.2%). Notably, single-contact tasks (egg transfer, peg-in-hole) showed both faster completion and higher success, while multi-stage tasks (cube stacking, fruit placement, handover) showed higher success at the cost of increased time. We analyze this task-dependent pattern in Sec. IV-F.

Parallelism-Constraint Ablation. Table V reports a focused ablation on the simulation cube task using a subset of four participants (F–I). We compare default retargeting and retargeting with the fingertip parallelism constraint under otherwise identical conditions. The parallelism constraint improved blue-cube success in both conditions (w/o haptic: 35.0% $\rightarrow$ 50.0%, +15.0 points; w/ haptic: 55.0% $\rightarrow$ 75.0%, +20.0 points). For yellow-cube attempts, performance improved from 5.0% to 10.0% without haptics and remained unchanged at 5.0% with haptics. These results are consistent with the intended benefit of parallel fingertip contact alignment for contact-critical grasp stages.

Subjective Experience. Figure 6b shows subjective ratings on 5-point Likert scales (1=strongly disagree, 5=strongly agree) across three dimensions. In simulation, ease improved from 2.78 ± 0.97 to 3.44 ± 0.53 ($p < 0.05$), contact clarity improved from 2.78 ± 1.20 to 4.67 ± 0.71 ($p = 0.0117$), and grasp confidence improved from 2.78 ± 1.09 to 3.89 ± 0.78 ($p = 0.0278$). In real-world operation, ease

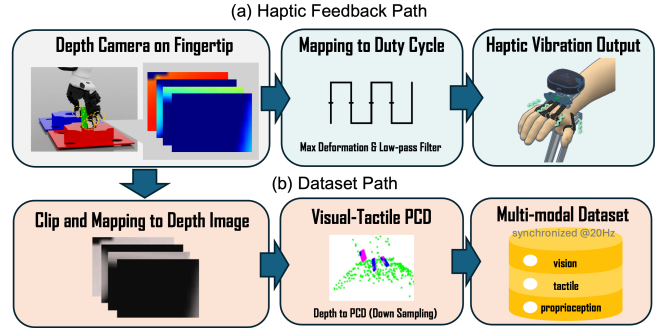


Fig. 7. Dual-purpose tactile proxy pipeline in simulation. A depth camera on each fingertip approximates 9DTact deformation. Top path (blue): the signal drives real-time haptic feedback to the operator via max-deformation extraction and PWM mapping. Bottom path (beige): the same depth data is converted to tactile point clouds and fused with visual observations into a synchronized multi-modal demonstration dataset for downstream policy learning.

changed from 3.33 ± 1.00 to 3.56 ± 0.88 ($p = 0.3173$, n.s.), contact clarity improved from 3.00 ± 1.12 to 4.22 ± 0.44 ($p = 0.0112$), and grasp confidence improved from 3.11 ± 1.17 to 4.00 ± 0.71 ($p = 0.0276$). Overall, the subjective results indicate consistent gains in contact perception and confidence.

E. Imitation Learning Validation

As shown in Fig. 7, the same depth-camera proxy that drives operator haptic feedback also provides tactile point clouds for policy learning, enabling a unified pipeline from data collection to training. We extend 3D-ViTac’s [30] visual-tactile representation to dexterous hands with multi-fingertip tactile sensing. The unified point cloud representation integrates visual and tactile information: each point contains (x, y, z) spatial coordinates, a one-hot encoding indicating point type (visual or tactile), and normalized force/deformation depth from tactile sensors. For dexterous hands, tactile point clouds from three fingertips are transformed to the global frame via forward kinematics and concatenated with visual point clouds from RGB-D cameras (Fig. 7, bottom path). The rest of the policy architecture (encoder and diffusion decoder) remains the same as 3D-ViTac.

Training Data. We collected approximately 50 demonstrations from three tasks: (c) peg-in-hole, (d) cube stacking,

TABLE VI
IMITATION LEARNING POLICY ROLLOUT RESULTS

Task	Vision	Visual-Tactile	Δ
Peg-in-hole	10%	27%	+17%
Cube stacking	27%	36%	+9%
Fruit placement	11%	9%	-2%
Average	16%	24%	+8%

Results from 100 rollout episodes per task.

and (e) fruit pick-and-place using ViHaTeleop. Each demonstration includes synchronized RGB-D images, 9DTact tactile data, and robot states and actions all recorded at 20 Hz.

Training Details. All policies were trained for 3,000 epochs. We report results using the final checkpoint for all methods to ensure fair comparison.

Baselines. We compare two conditions: (1) vision-only and (2) visual-tactile.

Results. Table VI shows rollout success rates across 100 evaluation episodes per task. Visual-tactile policies achieved 24% average success compared to 16% for vision-only (+8 percentage points), with task-dependent behavior. The largest improvement appears on peg-in-hole (10% $\rightarrow$ 27%, +17 points), cube stacking also improved (27% $\rightarrow$ 36%, +9 points), and fruit placement decreased slightly (11% $\rightarrow$ 9%, -2 points). These results suggest tactile cues are most helpful for contact-critical subtasks in this dataset, while additional tactile input may provide limited benefit for more vision-dominant tasks under the current data scale.

F. Discussion

The updated 9-participant study shows mixed objective outcomes and consistent subjective gains. Success rate increased across all six tasks, but completion time improved in only two tasks and increased in four. By contrast, subjective ratings showed statistically significant improvements in contact clarity and grasp confidence in both simulation and real-world settings. Within our tested protocol, these results support that low-cost vibrotactile feedback (<\$10 per hand) can improve teleoperation quality without motorized force-feedback hardware. The contrast between non-significant per-task objective tests and significant subjective results is consistent with statistical power: per-task comparisons use 5 trials per condition per participant, while subjective ratings aggregate each participant’s holistic experience across all trials.

Visual-Haptic Sensory Trade-offs. A consistent pattern in subjective data is that gains were generally larger in simulation than in real-world operation for contact clarity (+1.89 vs. +1.22) and grasp confidence (+1.11 vs. +0.89), while ease changes were smaller and not statistically significant in either setting. A plausible interpretation is that the simulation interface imposed stronger visual constraints, increasing the marginal utility of tactile cues. This interpretation is hypothesis-consistent with our observations, but it is not a controlled causal test of visual quality.

This suggests a practical design direction: in co-located settings with richer visual cues, haptics may play a complementary role; in constrained-view settings, haptics may provide larger relative gains. Dedicated experiments that

explicitly vary visual conditions are needed before making stronger substitution claims.

Speed-Accuracy Tradeoff and Task Structure. The completion-time results are consistent with the speed-accuracy tradeoff reported in haptic teleoperation [36], where richer sensory feedback encourages more deliberate manipulation strategies. However, the tradeoff is not uniform: single-contact tasks (egg transfer, peg-in-hole) showed *both* faster completion and higher success, while multi-stage tasks (cube stacking, fruit placement, handover) showed higher success at the cost of increased time. A plausible explanation is that for single-contact tasks, haptic feedback reduces exploratory retries by confirming contact onset on the first attempt, whereas for multi-stage tasks, operators process haptic cues at each contact transition, accumulating a time cost across stages even as each stage becomes more reliable. We also note that completion times are computed only over successful trials, so the non-haptic baseline selectively reflects fast attempts that were actually successful—the numerous quick failures are left out of the time comparison.

Imitation Learning Validation. The policy experiments indicate that ViHaTeleop can provide multimodal data suitable for downstream learning. The largest observed gain was on peg-in-hole (+17 points over vision-only), indicating that tactile information can improve contact-critical tasks in our setup.

The task-dependent patterns reflect modality trade-offs: tactile benefits contact-critical subtasks (insertion, stacking) but, with limited training data, may not improve vision-driven tasks (fruit localization). These results provide preliminary support for the pipeline from haptic teleoperation to multimodal policy learning, while highlighting open questions about selective modality fusion and data-efficient multimodal learning. The simulation environments, tactile proxy, and demonstration datasets will be released to support future research in visual-tactile dexterous manipulation.

Limitations. We evaluate through controlled ablations within our system architecture rather than cross-system comparisons, which would introduce multiple confounding variables. Future work could explore comparative benchmarking across different teleoperation paradigms and systematically vary visual feedback quality to further characterize the visual-haptic trade-off. We also note three remaining gaps for this revision cycle: no dedicated tracker/camera absolute-accuracy benchmark, no isolated quantitative ablation for LED/retargeting modules, and a limited participant pool size ($n = 9$).

V. CONCLUSIONS

We presented ViHaTeleop, a low-cost (\$550), lightweight (0.7 kg) teleoperation system that integrates haptic feedback into a vision-based platform optimized for contact-critical manipulation data collection.

Our user study (Sec. IV-C) with nine participants across six contact-critical tasks showed that haptic feedback increased success rates across all tasks (+2.2 to +15.6 points), while completion-time effects were task-dependent (Tables III, IV). Subjective ratings (Fig. 6) showed significant

improvements in contact clarity and grasp confidence in both simulation and real-world settings, with non-significant changes in ease.

The haptic integration pipeline—consisting of tactile-aware retargeting (Fig. 3), multi-fingertip 9DTact sensing, and LRA feedback mapping (Sec. III-E)—combined with targeted optimizations (fisheye camera for reduced torque, LED lighting for illumination robustness) achieves a favorable balance across cost, weight, and performance for large-scale data collection.

Our imitation learning experiments (Sec. IV-E) provided preliminary pipeline validation from teleoperation data collection to policy execution. Visual-tactile policies trained on ViHaTeleop demonstrations achieved 24% average success versus 16% for vision-only (Table VI), indicating potential benefits of tactile information for contact-critical tasks in this dataset.

The simulation-integrated pipeline—including task environments, a depth-camera tactile proxy, and multi-fingertip point cloud registration—provides reusable infrastructure for future visual-tactile research with dexterous hands.

We are committed to open-sourcing the implementation upon acceptance to lower barriers for dexterous manipulation research. Future work will scale to larger datasets and extend to diverse robot morphologies.

REFERENCES

- [1] A. O'Neill, *et al.*, “Open X-Embodiment: Robotic learning datasets and RT-X models,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024.
- [2] A. Khazatsky, *et al.*, “Droid: A large-scale in-the-wild robot manipulation dataset,” arXiv preprint arXiv:2403.12945, 2024.
- [3] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, X. He, X. Huang, *et al.*, “Agibot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems,” in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2025.
- [4] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, “Bridge Data: Boosting generalization of robotic skills with cross-domain datasets,” in *Proc. Robotics: Science and Systems (RSS)*, 2022.
- [5] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, “BridgeData V2: A dataset for robot learning at scale,” in *Proc. Conf. on Robot Learning (CoRL)*, 2023.
- [6] R. Zhong, C. Cheng, J. Xu, Y. Wei, C. Guo, D. Zhang, W. Dai, and H. Lu, “NuExo: A Wearable Exoskeleton Covering all Upper Limb ROM for Outdoor Data Collection and Teleoperation of Humanoid Robots,” arXiv preprint arXiv:2503.10554, Mar. 2025.
- [7] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang, “Bunny-VisionPro: Real-Time Bimanual Dexterous Teleoperation for Imitation Learning,” arXiv preprint arXiv:2407.03162, Jul. 2024.
- [8] HTC Corporation, “VIVE Ultimate Tracker,” <https://www.vive.com/us/accessory/vive-ultimate-tracker/>, Accessed: 2025.
- [9] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, “ACE: A Cross-platform Visual-Exoskeletons System for Low-Cost Dexterous Teleoperation,” in *Proc. Conf. on Robot Learning (CoRL)*, 2024, pp. 4895–4911.
- [10] Franka Emika GmbH, “Franka Research 3,” <https://franka.de/research>, 2024.
- [11] K. Shaw, A. Agarwal, and D. Pathak, “LEAP Hand: Low-Cost, Efficient, and Anthropomorphic Hand for Robot Learning,” in *Proc. Robotics: Science and Systems (RSS)*, 2023.
- [12] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, “AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System,” arXiv preprint arXiv:2307.04577, Jul. 2023.
- [13] C. Lin, *et al.*, “9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation,” *IEEE Robot. Autom. Lett.*, vol. 9, no. 2, pp. 923–930, 2023.
- [14] NVIDIA Corporation, “Isaac Sim: Robotics Simulation and Synthetic Data Generation,” <https://developer.nvidia.com/isaac-sim>, 2024.
- [15] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. Karen Liu, “DexCap: Scalable and Portable Mocap Data Collection System for Dexterous Manipulation,” in *Proc. Robotics: Science and Systems (RSS)*, July 2024.
- [16] A. Handa, *et al.*, “Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2020, pp. 9164–9170.
- [17] Y. Park and P. Agrawal, “Using Apple Vision Pro to Train and Control Robots,” 2024. [Online]. Available: <https://github.com/Improbable-AI/VisionProTeleop>
- [18] Q. Liu, Q. Ye, Z. Sun, Y. Cui, G. Li, and J. Chen, “Masked Visual-Tactile Pre-training for Robot Manipulation,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 13859–13875.
- [19] L. Heng, *et al.*, “ViTacFormer: Learning Cross-Modal Representation for Visuo-Tactile Dexterous Manipulation,” arXiv preprint arXiv:2506.15953, 2025.
- [20] M. Xu, *et al.*, “DexUMI: Using Human Hand as the Universal Manipulation Interface for Dexterous Manipulation,” arXiv preprint arXiv:2505.21864, 2025.
- [21] H.-S. Fang, *et al.*, “DEXOP: A Device for Robotic Transfer of Dexterous Human Manipulation,” arXiv preprint arXiv:2509.04441, 2025.
- [22] Y. Rong, T. Shiratori, and H. Joo, “Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration,” arXiv preprint arXiv:2008.08324, 2020.
- [23] F. Zhang, *et al.*, “Mediapipe hands: On-device real-time hand tracking,” arXiv preprint arXiv:2006.10214, 2020.
- [24] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware,” arXiv preprint arXiv:2304.13705, Apr. 2023.
- [25] P. Wu, *et al.*, “GELLO: A General, Low-Cost, and Intuitive Teleoperation Framework for Robot Manipulators,” in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2024.
- [26] H. Fang, *et al.*, “AirExo: Low-Cost Exoskeletons for Learning Whole-Arm Manipulation in the Wild,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 15031–15038.
- [27] H. Fang, *et al.*, “AirExo-2: Scaling up Generalizable Robotic Imitation Learning with Low-Cost Exoskeletons,” arXiv preprint arXiv:2503.03081, 2025.
- [28] J. Kulozik and N. Jarrassé, “Evaluating the precision of the HTC VIVE Ultimate Tracker with robotic and human movements under varied environmental conditions,” arXiv preprint arXiv:2409.01947, 2024.
- [29] H. Xue, *et al.*, “Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation,” in *Proc. Robotics: Science and Systems (RSS)*, 2025.
- [30] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, “3D-ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing,” in *Proc. Conf. on Robot Learning (CoRL)*, 2024.
- [31] H. Zhang, *et al.*, “DOGlove: Dexterous Manipulation with a Low-Cost Open-Source Haptic Force Feedback Glove,” arXiv preprint arXiv:2502.07730, 2025.
- [32] I. Sarakoglou, *et al.*, “Hexotrac: A highly under-actuated hand exoskeleton for finger tracking and force feedback,” in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2016.
- [33] J. Hummel, *et al.*, “A lightweight electrotactile feedback device for grasp improvement in immersive virtual environments,” in *Proc. IEEE Virtual Reality (VR)*, 2016. [Online]. Available: <http://dx.doi.org/10.1109/VR.2016.7504686>
- [34] W. Lin, *et al.*, “Super-resolution wearable electrotactile rendering system,” *Science Advances*, vol. 8, no. 36, Sep. 2022. [Online]. Available: <http://dx.doi.org/10.1126/sciadv.abp8738>
- [35] W. Yuan, *et al.*, “Gelsight: High-resolution robot tactile sensors for estimating geometry and force,” *Sensors*, vol. 17, no. 12, 2017, Art. no. 2762.
- [36] M. C. Yip, M. Tavakoli, and R. D. Howe, “Performance analysis of a haptic telemanipulation task under time delay,” *Advanced Robotics*, vol. 25, no. 5, pp. 651–673, 2011.