Contact Imitation Interface: Transferring Dexterous Skills by Showing How Contact Unfolds

Anonymous Authors

***Abstract*— Robot-free human demonstrations offer a promising route to scaling dexterous manipulation data. The key challenge is to preserve the demonstrator's natural dexterity while producing contact-grounded observations and actions that transfer to robot hands. We present Contact Imitation Interface (CIMI), a framework with contact-grounded demonstration and learning interfaces. The demonstration interface combines shared contact geometry and fingertip RGB views with contact-alignment-first retargeting and reprojection of wider scene views. The learning interface introduces a hierarchical policy in which a contact-surface planner predicts relative contact-surface trajectories and a controller coordinates wrist and finger commands to improve contact alignment. CIMI combines contact-grounded observations, retargeting, and action representations, improving mean final-stage success by at least 50 percentage points over baseline policies across six hand–task pairs on LEAP and Wuji. Both the hardware designs and software will be publicly available.**

I. Introduction

Low-cost teleoperation and portable hand-held grippers have enabled scalable demonstration collection for parallel-jaw manipulation [1, 2]. A human can fully control a gripper's low-dimensional action space through its pose and opening, using either the robot itself or a portable gripper to demonstrate manipulation. Extending this approach to dexterous hands is especially challenging for contact-sensitive tasks: differences in human and robot kinematics and contact-surface geometry make it difficult to reproduce the demonstrated sequence of contacts with the object. This calls for a demonstration interface that preserves natural human dexterity while capturing contact-grounded observations and motion that transfer to robot hands.

Existing approaches anchor demonstrations in either the human or robot embodiment. Human-centric interfaces prioritize natural manipulation with minimal constraint. Video and motion capture recover human hand motion [3–6], while instrumented gloves and wearable devices can additionally capture contact geometry or tactile interaction [7–10]. The recorded hand motion and sensory observations remain tied to the human embodiment: retargeting may not preserve the demonstrated contact evolution, and recorded observations may differ from those available during robot execution. Robot-centric interfaces instead establish compatibility through the collection hardware. Teleoperation records robot-native observations and actions [1, 11, 12], while exoskeletons and linkages mechanically map human motion to robot configurations [13–15]. This compatibility often requires robot-specific hardware or operator expertise and can restrict demonstration motion. The resulting trade-off is between demonstrator freedom and robot-side fidelity (Table I).

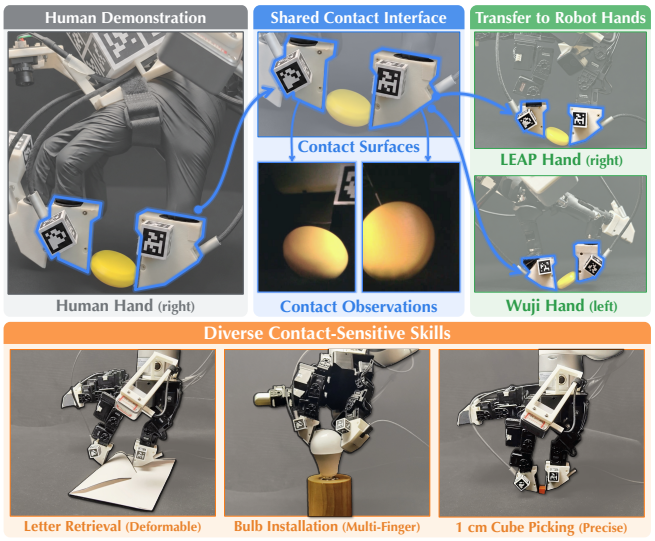

Fig. 1. **CIMI transfers contact-sensitive skills from human demonstrations to robot hands.** Shared fingertip modules provide corresponding contact surfaces and camera views for human demonstration and execution across different robot hands. Bottom: tasks involving deformable objects, multi-finger coordination, and precise contact.

To address this trade-off, we introduce the **Contact Imitation Interface (CIMI)**, which combines contact-grounded demonstration and learning interfaces:

1) We design a contact-grounded demonstration interface. Human and robot hands use fingertip modules with matching contact-surface geometry and camera placement. Each module includes an RGB camera, termed a *contact camera*, that captures local fingertip–object interactions. Cameras mounted at the wrist and thumb–index web space (*context cameras*) provide wider scene views for manipulation. Transferring demonstrations requires aligning contact motion and visual observations between human and robot hands. However, differences in hand kinematics can prevent simultaneous alignment of contact surfaces and context-camera poses. We therefore propose contact-alignment-first retargeting and context-view reprojection. By jointly optimizing wrist and finger motion, retargeting prioritizes reproducing the demonstrated contact; we retain the contact-camera images and reproject context views to the robot viewpoints.

2) We develop a contact-grounded learning interface. Policies commonly predict relative wrist motion and finger-joint commands [13]. In contact-sensitive tasks with high-degree-of-freedom hands, we observe that this representation can lead to poor wrist–finger coordination and contact misalignment. We propose a hierarchical policy where a contact-surface planner predicts relative contact-surface trajectories

Interface	Focus	Demonstrator dexterity		Demonstration–execution interface			Cross-hand reuse
		Motion freedom	Haptic feedback	Contact geometry	Contact observations	Action mapping	
Bare-hand video [5, 6, 16–22]	Human	Unconstrained	Direct touch	Mismatched human contact surfaces	None	Hand-pose retargeting	✓
Sensing glove [7–9, 23]	Human	Near-free	Through glove	Glove contact surfaces; matching varies	Human/shared tactile (optional)	Glove-state retargeting	✓
Teleoperation [11, 12, 24–33]	Robot	Controller/ robot-limited	Remote/optional	Matched robot contact surfaces	Robot sensing (optional)	Native robot commands	✓/×
Fingertip exoskeleton [13, 15, 34]	Robot	Exoskeleton-limited	Through contact surfaces	Matched robot contact surfaces	Matched tactile (optional)	Robot-aligned device state	×
Passive linkage [14, 35]	Robot	Linkage-limited	Through linkage	Robot-aligned contact surfaces	Robot-aligned tactile (optional)	Direct joint transfer	×
CIMI (Ours)	**Contact**	**Near-free**	**Through contact surfaces**	**Matched shared contact surfaces**	**Matched contact cameras**	**Contact-surface retargeting**	✓

TABLE I: Demonstration interfaces: dexterity, contact geometry and observations relative to execution, and cross-hand reuse. Mixed entries vary across cited systems.

from visual and proprioceptive inputs. A contact-surface controller converts these trajectories into coordinated wrist and finger commands.

By centering both demonstration and learning on contact surfaces, CIMI enables contact-sensitive dexterous skills to be learned from natural human demonstrations and reproduced across different robot hands. Experiments on LEAP and Wuji cover contact-sensitive tasks including grasping a 1 cm cube, wedging and lifting a flat coin, retrieving a letter, and screwing in a bulb. Across six hand–task pairs, CIMI improves mean success by at least 50 percentage points over baseline policies. Supplementary materials and videos are available at https://contactimitationinterface.github.io/.

II. Related Work

A. Demonstration Interfaces for Dexterous Manipulation

Human-centric interfaces capture manipulation directly from the human hand. Bare-hand approaches capture or reconstruct human-hand motion—and, when available, object motion—from RGB or egocentric videos [5, 6, 16, 18–21]. Motion-capture systems provide more reliable hand and wrist trajectories [3, 4], while instrumented gloves additionally capture touch, tactile signals, or contact-surface trajectories [7–9, 23]. For example, DexCap [3] combines hand motion capture with scene point clouds and fingertip-position retargeting. CIMI provides contact-grounded observations from matched fingertip cameras. Its contact-alignment-first retargeting improves contact alignment and task success over DexCap-style retargeting (Table III).

Robot-centric interfaces instead collect demonstrations through robot-compatible hardware. Teleoperation maps human input to an actuated robot hand and directly records robot actions and observations [11, 12, 24–32, 36, 37]. Fingertip exoskeletons constrain or measure human motion in a robot-oriented space [13, 34], while linkage-based exoskeletons and proxy hands mechanically align human motion with robot kinematics [14, 35]. For example, DexUMI [13] maps exoskeleton motion directly to robot joints, constraining demonstrations to the target hand's kinematics: 1–2 active DoFs per finger on Inspire and 2–3 on XHand. CIMI supports four active DoFs per finger without imposing robot-specific finger kinematics on the demonstrator. Its contact-grounded planner and controller improve success over direct wrist–joint prediction on our contact-sensitive tasks (Table II).

B. Learning Dexterous Manipulation from Human Demonstrations

We focus on learning robot policies from demonstrations captured through human hand motion and sensing, rather than transferring fully specified hand–object motion-capture trajectories [38, 39]. The key challenge is aligning human demonstrations with robot execution in both action and observation spaces. Existing methods address this challenge through explicit transformation, cross-embodiment policy learning, or both. For action alignment, human wrist, hand, or fingertip motion is retargeted into robot joint or end-effector commands [3, 5, 6, 9, 13, 40]. For observation alignment, demonstration videos can be transformed through human-hand segmentation, background inpainting, and robot-hand compositing, or matched sensing modalities can be used during both demonstration and execution [8, 9, 13]. Learning-based approaches use pretraining, co-training, or representation alignment to ground human-derived representations in robot sensing and control [4, 5, 18, 21].

CIMI uses matched fingertip contact surfaces as a shared cross-embodiment interface: tracked contact-surface trajectories supervise robot actions, while local sensing aligns policy observations, without imposing robot-specific hand kinematics on the demonstrator. A contact-surface planner and controller use this shared representation to separate contact-surface trajectory planning from wrist and finger control.

III. Contact Imitation Interface

CIMI uses contact surfaces to connect human demonstration and robot learning (Fig. 2). The demonstration interface transfers contact motion and visual observations to robot hands (Section III-A). The learning interface predicts relative contact-surface trajectories and converts them into coordinated wrist and finger commands (Section III-B).

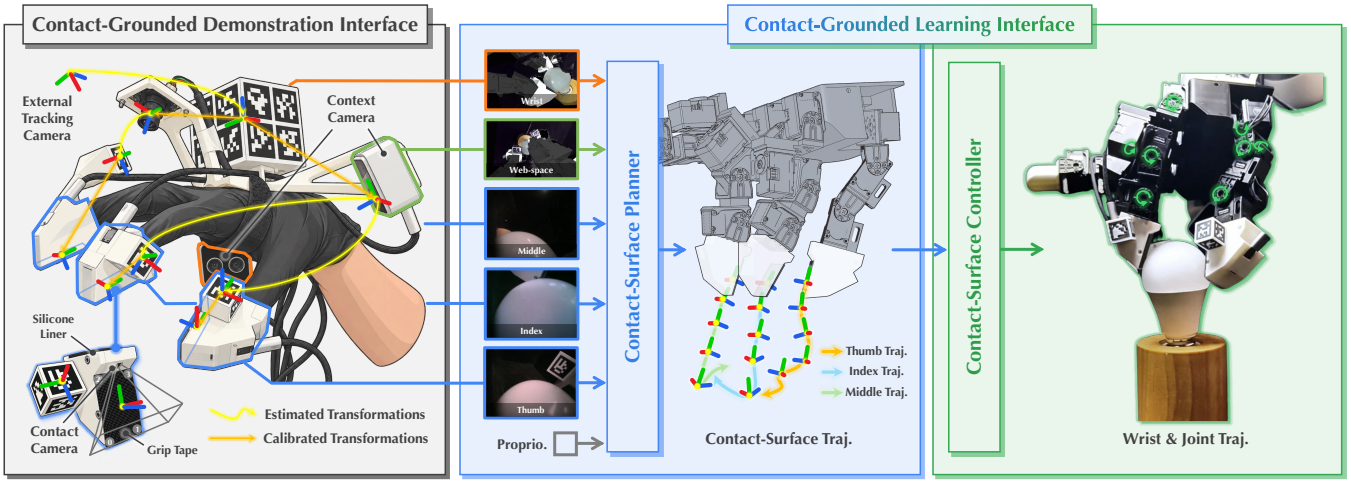

Fig. 2. **Contact connects demonstration and learning in CIMI.** The demonstration interface captures contact-surface trajectories and observations, then constructs robot demonstrations through contact-alignment-first retargeting and context-view reprojection. The learning interface predicts relative contact-surface trajectories and realizes them through coordinated wrist and finger control.

A. *Contact-Grounded Demonstration Interface*

The demonstration interface converts human demonstrations into contact-grounded robot motion targets and observations. We achieve this through four design choices. Shared fingertip hardware establishes matching contact-surface geometry and camera placement (**D1**), while visual tracking recovers demonstrated contact-surface trajectories (**D2**). Contact-alignment-first retargeting jointly optimizes wrist and finger motion to reproduce these trajectories (**D3**). We retain contact-camera views and reproject context-camera views to match the retargeted motion (**D4**).

D1. Contact-Grounded Capture Hardware. CIMI captures contact-surface trajectories and observations using cameras and printed fiducials. Human and robot fingertip modules share contact-surface geometry and camera placement. Each RGB camera (*contact camera*) views the local contact region from the fingertip root, tangentially to the surface. Two Gemini 305 stereo cameras view the wider scene (*context cameras*) from the wrist and thumb–index web space. We record nine synchronized streams at 10 Hz: three contact-camera views (thumb, index, and middle fingers), four context views, and middle-finger and external tracking views. Hardware specifications, fabrication details, and a bill of materials are provided in the supplementary material.

D2. Visual Contact-Surface Tracking. We recover contact-surface trajectories by composing estimated poses and calibrated transforms (Fig. 2). We use AprilTag markers [41] on AprilCube targets [42], with DeepTag [43] providing dense grid keypoints. We jointly fit a target pose to keypoints across visible tags in one multi-tag Perspective-n-Point (PnP) problem [44]. Visible tags provide redundant constraints when other tags are occluded. Frame definitions and estimator details are provided in the supplementary material.

D3. Contact-Alignment-First Retargeting. Our retargeting prioritizes contact-surface alignment over context-camera poses. Fixing the wrist trajectory to align human and robot context-camera poses (*context-alignment-first*) leaves only finger joints to reduce contact error (Fig. 3a). Following HoMMI's relaxation of camera-pose constraints [45], we free the wrist to improve contact alignment and correct the resulting viewpoint mismatch through reprojection (Fig. 3b; **D4**). Table III evaluates this choice. Human contact-surface poses and retargeted robot trajectories share the external tracking-camera frame. We optimize wrist poses $\mathbf{X}_t$ and finger joints $\mathbf{q}_t$ to match robot corners $\mathrm{FK}_{i,\ell}(\mathbf{X}_t, \mathbf{q}_t)$ to tracked human corners $\mathbf{p}^H_{i,\ell,t}$, where FK denotes forward kinematics. Matching four corners ℓ per finger i (Fig. 2, left) aligns contact-surface position and orientation:

$$\{\mathbf{X}^*_t, \mathbf{q}^*_t\}_{t=1}^T = \underset{\{\mathbf{X}_t, \mathbf{q}_t\}_{t=1}^T}{\operatorname{argmin}} \left[\sum_{t,i,\ell} \|\mathrm{FK}_{i,\ell}(\mathbf{X}_t, \mathbf{q}_t) - \mathbf{p}^H_{i,\ell,t}\|_2^2 + \mathcal{R} \right].$$

The smoothness regularizer $\mathcal{R}$ penalizes changes in wrist translation, rotation, and finger joints between valid neighboring frames. We initialize the optimization by aligning human and robot context-camera poses. We solve the retargeting optimization using PyRoki [46], subject to the robot's finger-joint limits. From the retargeted trajectory, we derive relative contact-surface trajectories to supervise the planner (Section III-B). The wrist and joint trajectories also provide the basis for motion augmentation and controller training. The retargeting objective and optimization settings are detailed in the supplementary material.

D4. Contact Observations and Context-View Reprojection. We construct demonstration observations to align with the contact and context views available during robot execution. Shared contact-surface geometry and camera placement give corresponding human and robot cameras the same nominal viewpoint relative to the local contact surface, so we retain the recorded contact-camera images. Context-camera viewpoints change with the retargeted wrist motion (**D3**), so we convert their recorded views through reprojection and robot-geometry overlay to match the robot perspective (Fig. 4). We optimize robot context-camera mounts using collected demonstrations (details in the supplementary material). For each stereo pair, we estimate depth with Fast-

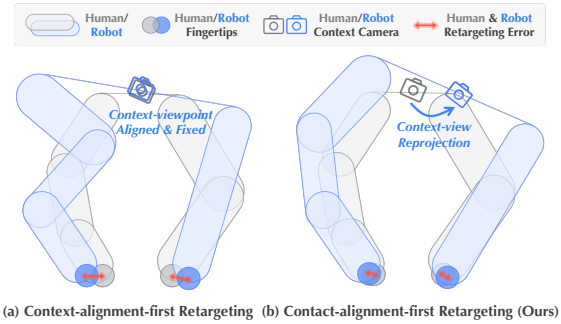

Fig. 3. **Retargeting priorities.** (a) Context-alignment-first retargeting matches context-camera poses but can leave contact surfaces misaligned. (b) Contact-alignment-first retargeting adjusts the wrist to improve contact alignment; reprojection corrects the viewpoint mismatch. Gray: human; blue: robot.

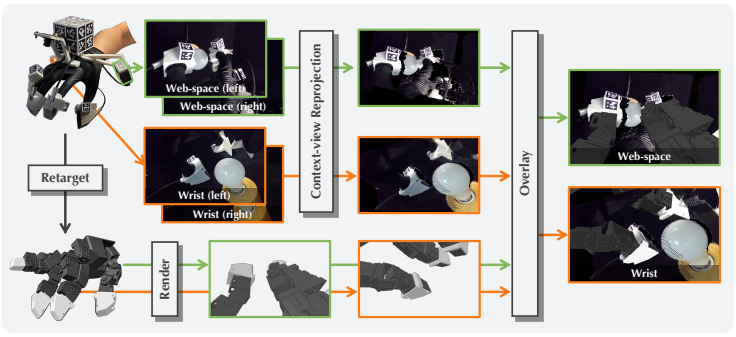

Fig. 4. **Context-view reprojection and robot-geometry overlay.** Recorded web-space (green) and wrist (orange) stereo images are reprojected to camera poses determined by retargeting. Robot geometry is rendered at the retargeted configuration and overlaid on these views to produce RGB training observations from the robot perspective. This conversion is performed offline.

FoundationStereo [47] and back-project the RGB images and estimated depth into a colored point cloud. We transform this cloud into the robot camera frame determined by these mounts and the retargeted wrist and hand configuration, reproject it onto the image plane, and composite it with the robot geometry using nearest-depth visibility. The resulting RGB context views, recorded contact-camera images, and retargeted wrist–hand states form synchronized demonstration samples for policy training. Demonstration conversion is performed offline. At execution, the policy receives contact RGB images and preprocessed context RGB views; depth is used only for context-image preprocessing.

B. Contact-Grounded Learning Interface

The learning interface separates contact-surface trajectory planning from wrist–finger control to improve coordination (**O3**). Direct command prediction [13] requires a visual policy to learn both contact motion and its kinematic realization. We instead represent actions as relative contact-surface trajectories (**L1**), train a visual planner to predict them from demonstrations (**L2**), and train a separate controller to convert them into wrist and finger commands using augmented synthetic motion data (**L3**).

L1. Relative Contact-Surface Trajectories as Actions. The planner predicts how each contact surface should move to establish and maintain contact with the object. We express this motion relative to the contact surface's current pose, aligning actions with observations from the contact camera on the same fingertip module.

Let F be the number of controlled fingers, and let $\mathrm{FK}_i(\mathbf{q}_t)$ denote the pose of contact surface i in the wrist frame. Its pose in a retargeted demonstration is $\mathbf{Y}_{i,t} = \mathbf{X}_t^* \mathrm{FK}_i(\mathbf{q}_t^*)$. For each contact surface, all future poses in a chunk use its frame at prediction time t as the reference, following the relative-trajectory convention of [2]:

$$\Delta\mathbf{T}_{i,t,h} = \mathbf{Y}_{i,t}^{-1}\mathbf{Y}_{i,t+h}.$$

Relative actions are invariant to fixed changes of the global frame's origin and orientation. Each pose uses three translation coordinates and six rotation coordinates (the rotation matrix's first two rows). Concatenating them across F fingers gives $\mathbf{u}_{t,h} \in \mathbb{R}^{9F}$.

L2. Contact-Surface Planner. A frozen C-RADIOv3-B backbone [48] extracts a summary token per contact or context view. View-specific projections and an MLP fuse the tokens into $\mathbf{z}_t^{\mathrm{vis}}$. The proprioceptive input $\mathbf{s}_t^{\mathrm{prop}}$ encodes thumb-relative contact-surface poses, exposing inter-finger geometry for contact coordination. This input outperforms the tested alternatives in Table IV. Given $\mathbf{o}_t = (\mathbf{z}_t^{\mathrm{vis}}, \mathbf{s}_t^{\mathrm{prop}})$, we train a flow-matching planner $\pi_\theta(\mathbf{U}_t \mid \mathbf{o}_t)$ [49] on demonstrated contact-surface trajectories. It outputs an H-step relative contact-surface trajectory $\widehat{\mathbf{U}}_t = [\widehat{\mathbf{u}}_{t,1}, \ldots, \widehat{\mathbf{u}}_{t,H}]$. We use visual augmentation during planner training to reduce sensitivity to reprojection artifacts, the appearance gap between rendered and real robot images, and photometric differences between demonstration and robot cameras. The supplement provides training details.

L3. Contact-Surface Controller. The MLP controller C_ϕ maps current joint values $\mathbf{q}_t$ and the predicted contact-surface trajectory $\widehat{\mathbf{U}}_t$ to future wrist poses $\widehat{\Delta\mathbf{X}}_{t,h}$ relative to the current wrist pose and absolute finger-joint targets $\widehat{\mathbf{q}}_{t,h}$:

$$\left[(\widehat{\Delta\mathbf{X}}_{t,h}, \widehat{\mathbf{q}}_{t,h})\right]_{h=1}^{H} = C_\phi(\mathbf{q}_t, \widehat{\mathbf{U}}_t).$$

To cover predicted contact-surface trajectories beyond recorded demonstrations, we augment retargeted trajectories from **D3** with smooth wrist and joint perturbations. Forward kinematics provides paired contact-surface trajectories and wrist–joint commands. We supervise the controller with wrist–joint command regression and contact-surface alignment measured through forward kinematics. Input construction, augmentation, and loss details are provided in the supplementary material.

At execution, the controller converts the trajectory into commands in one forward pass. We use $H = 16$ and execute an 8-step prefix, then replan from the latest observation. We reuse the controller across tasks with the same robot hand and controlled fingers.

IV. Experiments

We evaluate whether CIMI's contact-grounded demonstration and learning interfaces enable robot hands to

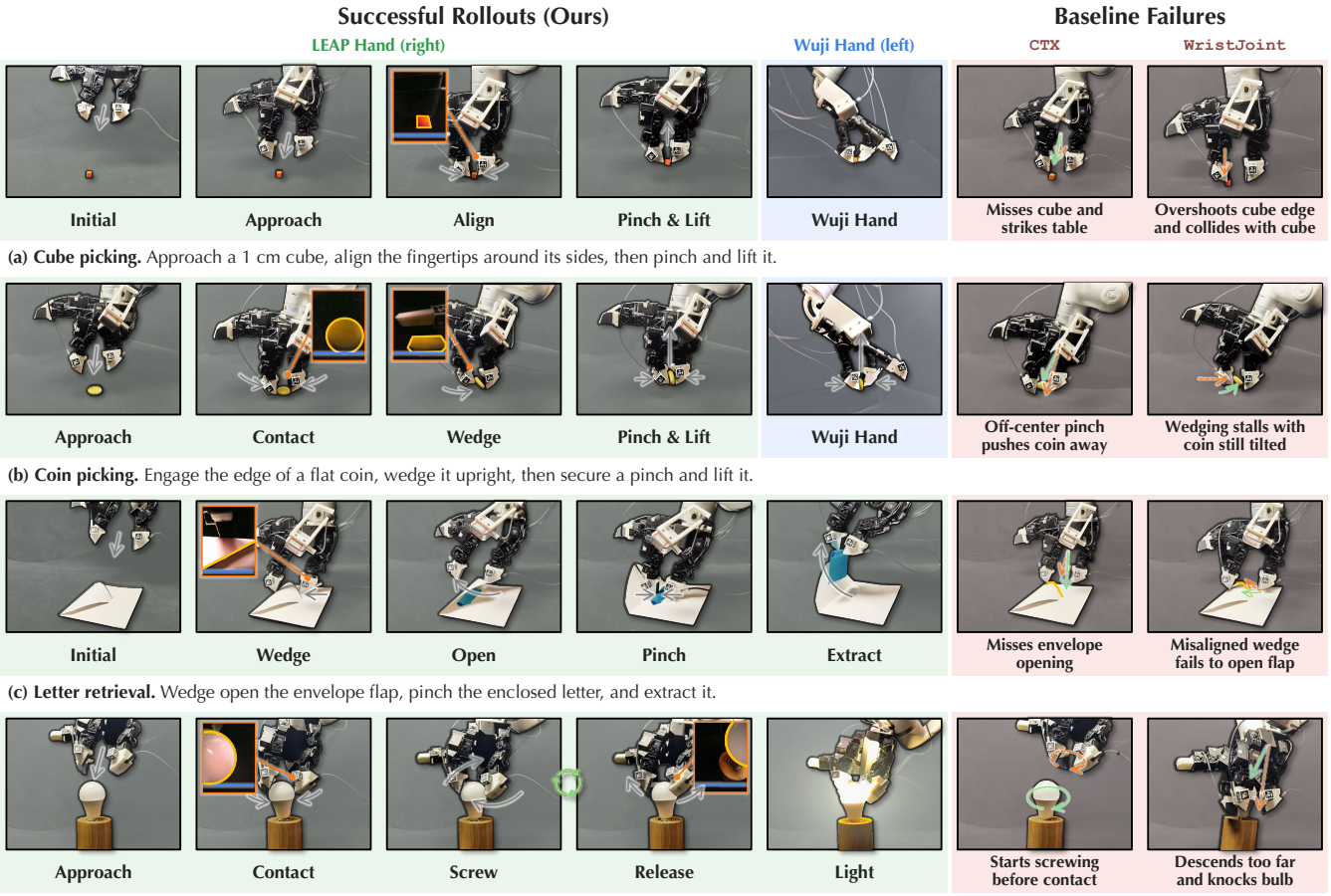

Fig. 5. **Real-world rollouts and failure cases.** Orange-bordered insets (▭) show contact-camera views; within each inset, blue lines (—) mark the contact surface and yellow outlines (—) mark object boundaries. Green (▭) and blue (▭) backgrounds show successful LEAP and Wuji rollouts, respectively; red (▭) marks `CTX` and `WristJoint` failures.

Policy	Visual	LEAP Hand								Wuji Hand				Mean
		Cube		Coin		Letter		Bulb		Cube		Coin		
		Align	Lift	Wedge	Lift	Open	Lift	Screw	Light	Align	Lift	Wedge	Lift	
`WristJoint`	`CTX`	0	0	0	0	0	0	35	5	0	0	0	0	0.8
`ContactSurface`	`CTX`	0	0	0	0	0	0	50	10	0	0	0	0	1.7
`WristJoint`	`CTX+CC`	35	20	50	35	25	15	85	70	30	15	40	30	30.8
`ContactSurface`	`CTX+CC`	**85**	**75**	**95**	**85**	**90**	**80**	**95**	**90**	**75**	**70**	**90**	**85**	**80.8**

TABLE II: Policy and observation comparison on LEAP and Wuji hands. SR in %; 20 rollouts per method, hand, and task. Mean SR averages final-stage SR equally over all six hand–task pairs.

learn contact-sensitive dexterous manipulation from human demonstrations. The experiments answer six questions: **Q1: Learning across robot hands.** Can CIMI enable different robot hands to learn contact-sensitive skills from the same human demonstrations? **Q2: Contact-grounded observations.** Do contact-grounded observations improve manipulation performance over context vision alone? **Q3: Contact-grounded policy.** Do hierarchical contact-surface planning and control improve manipulation over direct wrist and joint prediction? **Q4: Contact-surface alignment.** Is accurate contact-surface alignment important for contact-sensitive manipulation? **Q5: Collection efficiency.** How fast can CIMI collect demonstrations compared with bare-hand demonstration and teleoperation? **Q6: Design ablations.** Which design choices contribute most to reliable manipulation with CIMI?

A. Q1–Q3: Contact-Grounded Policy Evaluation

Tasks and Evaluation Protocol. Fig. 5 shows LEAP and Wuji rollouts. LEAP tasks comprise cube picking (146 training demos), coin picking (187 demos), letter retrieval (183 demos), and bulb installation (103 demos); Wuji tasks comprise cube picking (same 146 demos) and coin picking (same 187 demos). These tasks test fingertip alignment, wedging, envelope opening, and repeated bulb regrasping. We report stage-wise success rates (SR) over 20 rollouts

per method, hand, and task; mean SR averages final-stage success across the six hand–task pairs (Table II).

Compared Policies. Table II crosses two visual inputs—**`CTX`** (task-specific context-camera views) and **`CTX+CC`** (the same context views plus contact-camera images)—with two policies: **`WristJoint`** (direct prediction of relative wrist motion $(\mathbf{X}_t^*)^{-1}\mathbf{X}_{t+h}^*$ and joint changes $\mathbf{q}_{t+h}^* - \mathbf{q}_t^*$, following DexUMI [13]) and **`ContactSurface`** (our contact-grounded policy). Context views include wrist RGB and, depending on the task, web-space RGB; they are matched between **`CTX`** and **`CTX+CC`**. The supplementary material specifies the camera views for each task. All variants for a given hand use the same demonstrations and retargeted trajectories. Within each visual setting, policy comparisons match visual and proprioceptive inputs, flow-model width, and planner training budget. For cube and coin picking, we retarget the same human recordings to each hand and train separate policies.

O1: Human demonstrations support learning on both hands. CIMI achieves 80.8% mean final-stage SR across all six hand–task pairs with **`ContactSurface`** and **`CTX+CC`** (Table II). For the two tasks shared across hands, cube lifting succeeds in 75% of LEAP rollouts and 70% of Wuji rollouts, while coin lifting reaches 85% on both. LEAP also achieves 80% on letter retrieval and 90% on bulb installation. The successful rollouts in Fig. 5 show policies learned through CIMI reproducing demonstrated contact sequences on different robot hands.

O2: Contact-camera feedback improves contact establishment. Adding contact-camera views raises mean final-stage SR from 0.8% to 30.8% for **`WristJoint`** and from 1.7% to 80.8% for **`ContactSurface`** (Table II). With context vision alone, both policies score 0% even at the first stage of cube, coin, and letter tasks. The context-only examples in Fig. 5 show errors at contact establishment: missing the cube and striking the table, pushing the coin away with an off-center pinch, missing the envelope opening, and starting to screw before contacting the bulb. These examples illustrate the local placement and engagement cues that contact-camera views provide beyond context vision.

O3: Contact-surface planning better uses contact feedback. With **`CTX+CC`**, **`ContactSurface`** exceeds **`WristJoint`** by 50.0 percentage points in mean final-stage SR (80.8% versus 30.8%). Gains span all six hand–task pairs, ranging from 20 points on LEAP bulb to 65 on LEAP letter, with 55-point gains on both Wuji tasks. The **`WristJoint`** examples in Fig. 5 show residual motion errors: overshooting the cube edge, stalling while wedging the coin, misaligning with the envelope flap, and descending too far onto the bulb. This pattern supports the design in Section III-B: plan motion in the contact-surface frame and use a separate controller to coordinate wrist and finger commands. The comparison evaluates this combined design; Q6 examines individual input and augmentation choices.

Retargeting		Coin			Bulb		
Target	Priority	Error (mm)↓	SR (%)↑		Error (mm)↓	SR (%)↑	
			Wedge	Lift		Screw	Light
Centers	Context	15.37	0	0	7.85	10	0
Corners	Context	14.67	0	0	5.72	25	0
Centers	Contact	9.31	0	0	4.36	50	20
Corners	**Contact**	**0.51**	**95**	**85**	**1.09**	**95**	**90**

TABLE III: Retargeting ablation on LEAP. Priority denotes context- or contact-alignment-first. Gray: DexCap-style [3]; blue: ours. Error is the mean contact-corner distance over valid frames. SR uses 20 rollouts per variant and task; our SRs are from Table II.

B. Q4: Contact-Surface Alignment

Tasks and Evaluation Protocol. We compare center and corner matching under context- and contact-alignment-first retargeting on LEAP coin picking and bulb installation (Table III). Center matching aligns the centroid of each surface's four corners without constraining orientation. *Context-alignment-first* fixes the wrist to match context-camera poses; *contact-alignment-first* jointly optimizes the wrist and finger joints to align contact surfaces. All variants use the same human recordings. We measure mean Euclidean error (mm) between retargeted and demonstrated surface corners (four per finger) (**D3**). For each task, we retrain planners for all variants with identical data splits, architectures, training budgets, and controllers, then evaluate each in 20 rollouts.

O4: Contact priority and corner matching jointly improve contact transfer. All three alternative retargeting variants yield 0% Coin Wedge SR but 10–50% Bulb Screw SR (Table III). This contrast may reflect greater tolerance to contact error during bulb engagement than during coin-edge wedging. Completing bulb installation remains sensitive to alignment: both context-alignment-first variants have 0% Light SR. With contact priority, replacing center matching with corner matching reduces bulb contact-corner error from 4.36 to 1.09 mm and raises Light SR from 20% to 90%.

C. Q5: Collection Efficiency

Tasks and Evaluation Protocol. We compare CIMI, bare hands, and teleoperation by counting successful demonstrations collected by the same operator in 15 minutes (Fig. 6). We evaluate this collection throughput (CT) on coin picking and bulb installation, using the same task success criteria across interfaces. Teleoperation combines Manus hand tracking, Vive wrist tracking, and GeoRT [36] hand retargeting. The timed protocol includes failed attempts, resets, and object randomization.

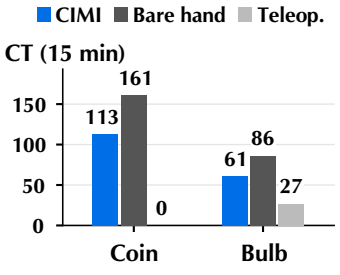

Fig. 6. **15-minute collection throughput on coin picking and bulb installation.**

O5: CIMI supports efficient collection of transferable demonstrations. CIMI retains 70% of bare-hand throughput

Variant	Coin		Bulb		Mean SR (%)↑
	Wedge	Lift	Screw	Light	
Context-view conversion					
No overlay	90	80	75	45	62.5
No reprojection, no overlay	75	50	55	10	30.0
Planner proprioceptive input					
No planner proprioception [13]	90	85	10	0	42.5
Joint values [1, 3]	70	55	80	50	52.5
Relative wrist + joints [2, 45]	60	40	30	5	22.5
Augmentation					
No planner visual augmentation	0	0	15	0	0.0
Task-only controller, no motion aug.	90	70	80	35	52.5
Ours	**95**	**85**	**95**	**90**	**87.5**

TABLE IV: Design ablations on LEAP. Stage-wise SR in %; 20 rollouts per variant and task. All variants use our `ContactSurface` + `CTX+CC` framework with the listed changes. Citations denote related input styles adapted to CIMI. Our SRs are from Table II. Mean SR averages Coin Lift and Bulb Light.

on coin picking and 71% on bulb installation (Fig. 6). It achieves 2.3× teleoperation throughput on bulb, while teleoperation yields no successful coin demonstrations. Together with Table II, these results support efficient demonstration collection for contact-sensitive robot learning.

D. Q6: Design Ablations

Tasks and Evaluation Protocol. We ablate context-view conversion, planner proprioception, and visual and motion augmentation on LEAP coin picking and bulb installation (Table IV). Each variant changes the listed components of `ContactSurface` with `CTX+CC`, using the same stage criteria and 20-rollout evaluation protocol.

O6: Context-view conversion improves contact control. Removing the robot-geometry overlay while retaining reprojection lowers mean SR from 87.5% to 62.5%; using the original human context views without either operation lowers it further to 30.0% (Table IV). In these rollouts, overall task motion remains reasonable, but contact control becomes less precise, despite unchanged contact-camera inputs. The observed errors suggest that reducing differences in context-camera viewpoints and robot appearance helps refine contact control.

O7: Thumb-relative proprioception improves planner performance. Thumb-relative contact-surface poses yield 87.5% mean SR, compared with 52.5% for joint-value inputs (style of [1, 3]). With joint values, coin rollouts show slightly looser fingertip contact, and Coin Lift SR falls from 85% to 55%. Removing planner proprioception (style of [13]) preserves 85% Coin Lift SR but reduces Bulb Light SR from 90% to 0%. In bulb rollouts, relative-pose errors accumulate over successive motions, and the middle finger drifts away from the bulb. The relative proprioception variant (style of [2, 45]) uses the wrist pose relative to the previous frame and inter-frame joint changes. This variant yields 22.5% mean SR and can stall during bulb screwing, predicting little or no motion.

O8: Visual augmentation and broader controller training improve success. Without planner visual augmentation, final-stage SR is 0% on both tasks, although Bulb Screw reaches 15%. Training the controller only on the target task's retargeted trajectories, without synthetic motion augmentation, reduces mean SR from 87.5% to 52.5%: Coin Lift falls from 85% to 70%, and Bulb Light from 90% to 35%. The larger drop on bulb installation suggests that this task benefits especially from broader controller training. This comparison changes both motion augmentation and training-data coverage.

V. Conclusion

CIMI connects natural human demonstrations to robot hands through contact-grounded observations, retargeting, and action representations. Across six hand–task pairs on LEAP and Wuji, it improves mean success by at least 50 percentage points over baseline policies. These results support shared contact surfaces as a practical interface for transferring contact-sensitive skills across different hand kinematics. We will release the hardware designs and software to support broader robot-free demonstration collection. Extending CIMI to bimanual and mobile manipulation offers a path toward transferring a wider range of everyday skills.

References

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems XIX*, 2023.

[2] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Robotics: Science and Systems*, 2024.

[3] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "DexCap: Scalable and portable mocap data collection system for dexterous manipulation," in *Robotics: Science and Systems*, 2024.

[4] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "DexWild: Dexterous human interactions for in-the-wild robot policies," in *Robotics: Science and Systems*, 2025.

[5] R. Zheng, D. Niu, Y. Xie, J. Wang, M. Xu, Y. Jiang, F. Castañeda, F. Hu, Y. L. Tan, L. Fu, T. Darrell, F. Huang, Y. Zhu, D. Xu, and L. Fan, "EgoScale: Scaling dexterous manipulation with diverse egocentric human data," *arXiv preprint arXiv:2602.16710*, 2026.

[6] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "DexMV: Imitation learning for dexterous manipulation from human videos," in *European Conference on Computer Vision*, 2022, pp. 570–587.

[7] C. Lin and D. Zhao, "ART-Glove: Articulated tactile glove for contact-grounded dexterous interaction capture," *arXiv preprint arXiv:2606.16370*, 2026.

[8] J. Yin, H. Qi, Y. Wi, S. Kundu, M. Lambeta, W. Yang, C. Wang, T. Wu, J. Malik, and T. Hellebrekers, "OSMO: Open-source tactile glove for human-to-robot skill transfer," *arXiv preprint arXiv:2512.08920*, 2025.

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

[10] P. Sathe, A. Schmitz, T. P. Tomo, S. Somlor, S. Funabashi, and S. Sugano, "FingerTac—an interchangeable and wearable tactile sensor for the fingertips of human and robot hands," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023, pp. 10 813–10 820.

[11] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system," in *Robotics: Science and Systems*, 2023.

[12] C. Xu, Y. Jiang, J. Huan, Y. Fu, H. Zhou, W. Yuan, J. Yu, W. Zhang, H. Yuan, and Z. Lu, "RealDexUMI: A wearable universal manipulation interface for dexterous robot learning," *arXiv preprint arXiv:2606.06033*, 2026.

[13] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. Song, "DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation," in *Proceedings of the 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 305. PMLR, 2025, pp. 437–459.

[14] H.-S. Fang, B. Romero, Y. Xie, A. Hu, B.-R. Huang, J. Alvarez, M. Kim, G. Margolis, K. Anbarasu, M. Tomizuka, E. Adelson, and P. Agrawal, "DEXOP: A device for robotic transfer of dexterous human manipulation," *arXiv preprint arXiv:2509.04441*, 2025.

[15] J. Du, J. Ren, Q. Yu, N. Zhang, Y. Deng, X. Wei, Y. Liu, G. Gu, and X. Zhu, "MILE: A mechanically isomorphic hand exoskeleton and visuotactile robotic hand for data collection in dexterous manipulation," *arXiv preprint arXiv:2512.00324*, 2025.

[16] K. Shaw, S. Bahl, and D. Pathak, "VideoDex: Learning dexterity from internet videos," in *Proceedings of the 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205. PMLR, 2023, pp. 654–665.

[17] Z. Chen, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "ViViDex: Learning vision-based dexterous manipulation from human videos," in *IEEE International Conference on Robotics and Automation*, 2025, pp. 3336–3343.

[18] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "EgoMimic: Scaling imitation learning via egocentric video," in *IEEE International Conference on Robotics and Automation*, 2025, pp. 13 226–13 233.

[19] B. Paliwal, H. Etukuru, W. Liang, P. Abbeel, N. M. M. Shafiullah, and J. Malik, "Do as I do: Dexterous manipulation data from everyday human videos," *arXiv preprint arXiv:2606.19333*, 2026.

[20] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "EgoDex: Learning dexterous manipulation from large-scale egocentric video," in *International Conference on Learning Representations*, 2026.

[21] R. Punamiya, S. Kareer, Z. Liu, J. Citron, R.-Z. Qiu, X. Cai, A. Gavryushin, J. Chen, D. Liconti, L. Y. Zhu *et al.*, "EgoVerse: An egocentric human dataset for robot learning from around the world," *arXiv preprint arXiv:2604.07607*, 2026.

[22] H. Chen, T. Dong, T. Wu, L. Wang, Y. Jangir, Y. Niu, Y. Ye, H. Bharadhwaj, Z. Erickson, and J. Ichnowski, "Dexterous manipulation policies from RGB human videos via 3D hand-object trajectory reconstruction," *arXiv preprint arXiv:2602.09013*, 2026.

[23] Y. R. Song, J. Li, R. Fu, D. Murphy, K. Zhou, R. Shiv, Y. Li, H. Xiong, C. E. Owens, Y. Du, Y. Luo, X. Cheng, A. Torralba, W. Matusik, and P. P. Liang, "OPENTOUCH: Bringing full-hand touch to real-world interaction," *arXiv preprint arXiv:2512.16842*, 2025.

[24] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *IEEE International Conference on Robotics and Automation*, 2020.

[25] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "OPEN TEACH: A versatile teleoperation system for robotic manipulation," in *Proceedings of the 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 2372–2395.

[26] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-TeleVision: Teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[27] 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*, 2024.

[28] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," in *IEEE International Conference on Robotics and Automation*, 2025, pp. 5637–5643.

[29] H. Zhang, S. Hu, Z. Yuan, and H. Xu, "DOGlove: Dexterous manipulation with a low-cost open-source haptic force feedback glove," *arXiv preprint arXiv:2502.07730*, 2025.

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

[31] J. Koh, H. Jung, N. Kim, W. Ko, and C. Nam, "DEX-Mouse: A low-cost portable and universal interface with force feedback for data collection of dexterous robotic hands," *arXiv preprint arXiv:2604.15013*, 2026.

[32] X. Chao, S. Mu, Y. Liu, S. Li, C. Lyu, X.-P. Zhang, and W. Ding, "Exo-ViHa: A cross-platform exoskeleton system with visual and haptic feedback for efficient dexterous skill learning," *arXiv preprint arXiv:2503.01543*, 2025.

[33] O. Rayyan, M. Gilles, and Y. Cui, "TeleDex: Accessible dexterous teleoperation," *arXiv preprint arXiv:2603.17065*, 2026.

[34] Z. Si, J. E. Chen, M. E. Karagozler, A. Bronars, J. Hutchinson, T. Lampe, N. Gileadi, T. Howell, S. Saliceti, L. Barczyk, I. O. Correa, T. Erez, M. Shridhar, M. F. Martins, K. Bousmalis, N. Heess, F. Nori, and M. Bauza, "ExoStart: Efficient learning for dexterous manipulation with sensorized exoskeleton demonstrations," *arXiv preprint arXiv:2506.11775*, 2025.

[35] A. Zhu, M. Zhu, B. J. Kim, J. V. S. H. Ramos, Y. Shi, Y. Wu, R. Dhar, F. Yang, R. Hou, H. Fang, Q. Wang, Y. Cui, and D. W. Hong, "DexEXO: A wearability-first dexterous exoskeleton for operator-agnostic demonstration and learning," *arXiv preprint arXiv:2603.17323*, 2026.

[36] Z.-H. Yin, C. Wang, L. Pineda, K. Bodduluri, T. Wu, P. Abbeel, and M. Mukadam, "Geometric retargeting: A principled, ultrafast neural hand retargeting algorithm," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025, pp. 17 376–17 382.

[37] Z.-H. Yin, C. Wang, L. Pineda, F. R. Hogan, C. K. Bodduluri, A. Sharma, P. Lancaster, I. Prasad, M. Kalakrishnan, J. Malik, M. Lambeta, T. Wu, P. Abbeel, and M. Mukadam, "DexterityGen: Foundation controller for unprecedented dexterity," in *Robotics: Science and Systems*, 2025.

[38] K. Li, P. Li, T. Liu, Y. Li, and S. Huang, "ManipTrans: Efficient dexterous bimanual manipulation transfer via residual learning," *arXiv preprint arXiv:2503.21860*, 2025.

[39] Y. Chen, C. Wang, Y. Yang, and C. K. Liu, "Object-centric dexterous manipulation from human motion data," in *Proceedings of the 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 3828–3846.

[40] G. Zhang, Q. Xu, H. Zhang, J. Ma, L. He, Y. Bao, Z. Ping, Z. Yuan, C. Lu, C. Yuan, T. Liang, X. Tian, M. Shao, F. Zhang, M. Ding, Y. Gao, H. Zhao, H. Zhao, and H. Xu, "UniDex: A robot foundation suite for universal dexterous hand control from egocentric human videos," *arXiv preprint arXiv:2603.22264*, 2026.

[41] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 3400–3407.

[42] Y. Park and P. Agrawal, "AprilCube: 3D-printable fiducial targets for reliable 6-DoF pose estimation," Software, 2026. [Online]. Available: https://github.com/younghyopark/aprilcube

[43] Z. Zhang, Y. Hu, G. Yu, and J. Dai, "DeepTag: A general framework for fiducial marker design and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2931–2944, 2023.

[44] Z. Xu, Y. Li, X. Wu, T. Qiu, and L. Shao, "FingerEye: Learning dexterous manipulation with continuous vision-tactile sensing," *arXiv preprint arXiv:2604.20689*, 2026.

[45] X. Xu, J. Park, H. Zhang, E. Cousineau, A. Bhat, J. Barreiros, D. Wang, J. Bohg, and S. Song, "HoMMI: Learning whole-body mobile manipulation from human demonstrations," *arXiv preprint arXiv:2603.03243*, 2026.

[46] C. M. Kim, B. Yi, H. Choi, Y. Ma, K. Goldberg, and A. Kanazawa, "PyRoki: A modular toolkit for robot kinematic optimization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025, pp. 1312–1319.

[47] B. Wen, S. Dewan, and S. Birchfield, "Fast-FoundationStereo: Real-time zero-shot stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 7513–7524.

[48] G. Heinrich, M. Ranzinger, H. Yin, Y. Lu, J. Kautz, A. Tao, B. Catanzaro, and P. Molchanov, "RADIOv2.5: Improved baselines for agglomerative vision foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 22 487–22 497. [Online]. Available: https://arxiv.org/abs/2412.07679

[49] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations*, 2023.