Dex-X: Learning Visual-Tactile Dexterous Manipulation From Human Videos with Simulated Interaction

Ruoqu Chen[1,2], Feixiang Ruan[1,3*], Liu Cao[1,2], Zihao Wang[1], Botian Xu[1*], Shiqin Tong[1*], Jiajun Liu[1,2*], Mingzhi Pei[1*], Chenyu Zhang[1,2], Wanli Xing[3], Kaifeng Zhang[3], Mengdi Xu[1,2]
[1]IIIS, Tsinghua University,[2]Shanghai Qi Zhi Institute,[3]Sharpa,[*]Work done during internship at Tsinghua University.

***Abstract*—Human videos are an abundant source of dexterous manipulation behaviors, but they lack tactile information that is crucial for contact-rich interaction. This raises a fundamental question: can robots learn deployable visual-tactile dexterous manipulation policies directly from human video demonstrations? We present DEX-X, a framework for learning visual-tactile dexterous manipulation from human videos through simulation. Our key insight is that simulation can serve as a tactile completion engine. Given monocular human demonstrations, DEX-X reconstructs hand-object interactions in simulation, where physically grounded contact dynamics provide tactile supervision unavailable in the original videos. Leveraging this recovered tactile information, we train visual-tactile dexterous manipulation policies and distill them into deployable policies operating on point-cloud observations and tactile sensing. We demonstrate zero-shot sim-to-real transfer on dexterous hand-arm platforms across diverse grasping and contact-rich tool-use tasks. The teacher policy achieves 65.9% average success across eight tasks in simulation, while the distilled visual-tactile policy achieves a 53% success rate on cube picking and 23% on the challenging real-world table-cleaning task, which is not achieved by existing general sim-to-real manipulation systems. Our results suggest that tactile completion is a critical ingredient for scalable dexterous manipulation learning and provide a path toward learning real-world robot skills from Internet-scale human video data. More visualizations are on the project website: https://dexx-code.github.io.**

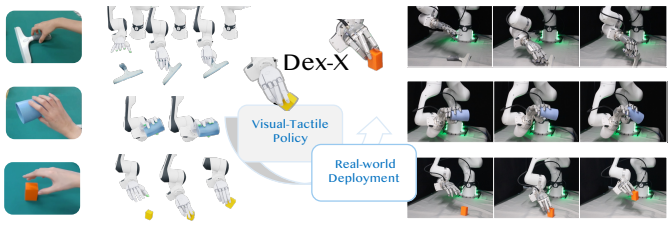

Fig. 1: **DEX-X Overview**. DEX-X learns deployable visual-tactile dexterous manipulation policies from human video demonstrations. It transfers human hand-object interactions into simulation, where tactile-aware reinforcement learning enriches visual demonstrations with contact information and learns deployable manipulation policies. The resulting policies transfer zero-shot to real-world grasping and tool-use tasks.

I. INTRODUCTION

Dexterous manipulation requires coordinating high-dimensional hand-arm motion under complex contact dynamics while continuously adapting to multimodal sensory feedback. Humans perform such behaviors effortlessly, seamlessly integrating vision and touch to grasp objects and execute contact-rich tool use across diverse environments. Replicating this visual-tactile dexterous manipulation capability on robotic platforms remains a long-standing challenge [1, 29, 30].

Recent reinforcement learning (RL) and imitation learning systems are increasingly capable [6, 31, 23, 14, 12], but both rely on large-scale data from simulation or specialized teleoperation [21], and collecting tactile or force demonstrations further demands instrumented gloves or force-feedback teleoperation hardware [13]—limiting scalability. Human videos are a compelling alternative: abundant, diverse, and captured in natural environments, they are arguably the most scalable source of manipulation data. This raises a fundamental question: *can robots learn deployable visual-tactile dexterous manipulation policies directly from human video demonstrations?*

This poses two challenges. First, videos lack tactile information—contact state, grasp stability, force, and slip—that is critical for contact-rich manipulation yet unobservable from vision alone. Second, transferring dexterous policies from simulation to hardware remains difficult due to perception noise, actuation delays, and contact-modeling inaccuracies.

We introduce DEX-X, which learns visual-tactile dexterous manipulation from human videos through simulation. Our key insight is that *simulation can serve as a tactile completion engine*: physically grounded interaction in simulation reconstructs the contact signals absent from videos, yielding paired visual-tactile trajectories. DEX-X learns contact-rich behaviors via tactile-aware RL and distills privileged experts into a multi-task policy over a unified point-cloud-and-contact input that transfers zero-shot to real hand-arm platforms. It reaches 65.9% average success across eight simulated tasks, and the distilled policy achieves 53% on real-world cube picking and 23% on the challenging table-cleaning task, outperforming baselines.

Our contributions are: (i) DEX-X, which uses simulation as a tactile completion engine to recover the contact information absent from human videos; (ii) a teacher-student paradigm pairing human-video motion priors with tactile-aware RL for

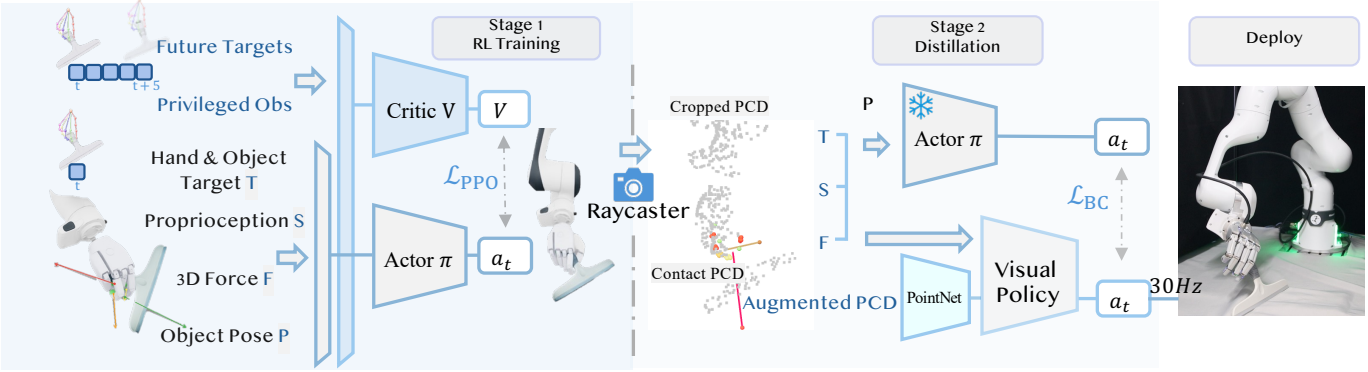

Fig. 2: **Training framework of DEX-X.** After transferring human demonstrations into simulation, we train a privileged state-based expert using RL with demonstration references, object states, and tactile contact information. The expert is distilled into a multi-task visual-tactile policy operating on a unified *contact point cloud* that embeds fingertip tactile feedback into the scene geometry and fuses vision, touch, and proprioception. The resulting policy transfers zero-shot to diverse real-world dexterous manipulation tasks.

contact-rich skills that transfer zero-shot; and (iii) zero-shot sim-to-real transfer on a 29-DoF hand-arm platform—among the first to perform real-world visual-tactile dexterous tool use learned directly from human video without task-specific fine-tuning.

II. RELATED WORK

Dexterous manipulation from human demonstrations. Prior work retargets human hand motion into robot embodiments [12, 14], learns object-centric skills from reconstructed motion [3], learns point-based policies from in-the-wild human video [5], or combines reconstruction with RL and visuomotor distillation [4]—but mostly uses vision-only supervision without modeling interaction forces. Physics-based retargeting recovers missing dynamics by sampling dynamically feasible trajectories [16]; we share this goal but, rather than generating trajectory data, learn closed-loop policies that embed recovered per-fingertip tactile feedback into a deployable visual-tactile representation. **Sim-to-real transfer.** Existing systems show agile in-hand skills [6, 28, 23] and object-level generalization via large-scale simulation [22, 30, 34, 10], but remain centered on grasp adaptation and object tracking, with tactile feedback underexplored. Others narrow the gap via residual real-world adaptation [26] or adaptive controller learning [33]; ours transfers zero-shot without real-world adaptation. **Visual-tactile manipulation.** Touch helps under occlusion [25, 27, 7, 8]; closest to us, Robot Synesthesia [32] augments point clouds with simulated hand-contact geometry for single-object in-hand rotation, and force-attending methods improve contact-rich policies but rely on real teleoperated force [13]. We instead add explicit per-fingertip force magnitudes to the point cloud and target hand-arm tool-object-environment interaction, recovering all tactile feedback from simulated contact with no physical tactile data [2].

III. DEX-X: LEARNING VISUAL-TACTILE MANIPULATION FROM HUMAN VIDEOS

DEX-X is a sim-to-real framework with three stages (Fig. 2): human hand-object interactions are reconstructed from video and retargeted into simulation (Sec. III-B); a privileged-state policy is trained via RL with tactile feedback and contact dynamics unavailable in the videos (Sec. III-C); and the policy is distilled into a deployable visual-tactile controller for zero-shot real-world transfer (Sec. III-D).

A. Problem Formulation

Given a dataset of monocular human demonstrations $\mathcal{D} = \{\tau_i^h\}_{i=1}^N$, where each trajectory $\tau_i^h = (I_0, I_1, \ldots)$ captures a human performing a manipulation task, our goal is a multi-task policy $\pi(a_t \mid o_t)$ mapping robot observations $o_t = \{o_t^{prop}, o_t^{vis}, o_t^{tac}\}$ (proprioception, vision, tactile) to joint-level commands a_t for both arm and hand. The central challenge is the mismatch between human demonstrations, which provide only vision, and the target policy, which depends on tactile feedback for reasoning about contact state, grasp stability, and object slip. The objective is to recover the missing tactile modality and learn a visual-tactile policy that performs robust contact-rich manipulation and transfers to real execution.

B. Motion Prior Extraction with Spatial Augmentation

Hand-object reconstruction. From a monocular demonstration τ^h, we reconstruct hand-object interaction trajectories at 30 Hz. Object poses are estimated with FoundationPose [24] and hand poses with WiLoR [18], refined via MANO-based temporal consistency and hand-object penetration constraints. The reference trajectory τ contains wrist motion, MANO keypoint trajectories, and object trajectory for reward computation.

Embodiment retargeting. We retarget the reconstructed human trajectory to the robot through multi-stage optimization: arm motion is initialized by inverse kinematics on the

wrist, followed by joint arm-hand optimization aligning robot fingertips with MANO targets (emphasizing thumb and index), with the resulting trajectories clamped to hardware joint limits to ensure feasibility.

Spatial augmentation. Following prior demonstration-augmentation pipelines [11, 9], we augment each demonstration before retargeting with a shared planar translation and yaw perturbation. For any planar point $\mathbf{x}_{xy}$ from the wrist, MANO target, or object trajectory, we apply $\tilde{\mathbf{x}}_{xy} = \mathbf{R}_z(\Delta\psi)(\mathbf{x}_{xy} - \mathbf{c}_{xy}) + \mathbf{c}_{xy} + \Delta\mathbf{p}_{xy}$, where $\mathbf{c}_{xy}$ is the arm-base position, $\Delta\psi \sim \mathcal{U}(-\psi_{\max}, \psi_{\max})$ is the yaw perturbation, and $\Delta\mathbf{p}_{xy}$ is a sampled planar translation within the target workspace.

C. State Expert Training

Kinematic retargeting provides motion references but is insufficient for contact-rich interaction. We leverage simulation to fill in per-fingertip contact forces, training a closed-loop state-based expert that uses the retargeted trajectories as behavioral priors while learning feedback corrections. This expert serves as an upper-bound policy and provides action supervision for distillation. We build on PPO [20] with an asymmetric actor-critic [17]; domain randomization covers object physics, PD gains, action delay, observation noise, and tactile sensing.

Tactile feedback from simulation. The expert receives five scalar per-fingertip contact-force magnitudes from simulated contact sensors, smoothed with an exponential moving average to suppress transient spikes, and randomized during training with latency, multiplicative noise, and per-finger dropout to improve robustness to real sensor artifacts.

Observation, action, and reward. The actor observes proprioception, the reference sequence τ_t (wrist pose, MANO keypoints, object pose, fingertip-distance features), a BPS geometry encoding [19], and tactile feedback; the critic additionally receives privileged simulation state and a 5-step future reference horizon. The policy outputs $\mathbf{a}_t \in R^{29}$ (7 arm joint-delta commands and 22 hand joint position targets), applied as $\mathbf{q}_{t+1}^{\text{arm}} = \mathbf{q}_t^{\text{arm}} + \alpha\,\mathbf{a}_t^{\text{arm}}$ with $\alpha = 0.2\,\text{rad}$ at 30 Hz, and low-pass filtered. The reward combines demonstration tracking with task completion,

$$r_t = w_w\, r_t^{\text{wrist}} + w_h\, r_t^{\text{hand}} + w_t\, r_t^{\text{task}} - w_a\, \mathbf{a}_t^2 - w_l\, r_t^{\text{limit}}, \quad (1)$$

where the task reward includes approach shaping, contact-stability bonuses, and a no-slip penalty on fingertip-object relative motion.

D. Policy Distillation and Sim-to-Real Transfer

To enable deployment, we distill the privileged expert into a visuomotor student $\pi_{\text{student}}(\mathbf{a} \mid \mathbf{o}^{\text{deploy}})$ via teacher-student imitation. The deployable observation consists of proprioception, depth-based point clouds, and tactile feedback, without access to demonstration trajectories or privileged object state; the student is optimized to match the teacher's action predictions under identical simulation rollouts, transferring interaction-aware control from privileged state space to partial observations.

Tactile-augmented point cloud. The student uses a unified point cloud $\mathcal{P}_t = \mathcal{P}_t^{\text{scene}} \cup \mathcal{P}_t^{\text{hand}} \cup \mathcal{P}_t^{\text{tac}}$, where $\mathcal{P}_t^{\text{scene}}$ is back-projected from the depth image, $\mathcal{P}_t^{\text{hand}}$ is generated from robot joint encoders via forward kinematics, and $\mathcal{P}_t^{\text{tac}}$ consists of fingertip surface points augmented with tactile force features. The three sets are concatenated with type embeddings and encoded by a shared PointNet backbone before action prediction. At deployment, the policy runs in a closed loop, constructing the multimodal observation from onboard depth, joint encoders, and tactile readings and executing a single forward pass for real-time control. We apply geometric perturbation on scene points, dropout on partial observations, and tactile-noise injection during training for robustness.

IV. Experiments

We structure evaluation around three questions: **Q1**: Can DEX-X handle various dexterous manipulation tasks? **Q2**: How do different sensory modalities and their representations contribute? **Q3**: Can the trained policy transfer to the real world?

A. Experimental Setup

We evaluate on a Franka FR3 arm with a Sharpa HA4 dexterous hand (29 DoF per arm). Visual observations come from a fixed RealSense depth camera and tactile feedback from fingertip sensors. We train in IsaacLab [15] using PPO with an asymmetric actor-critic across 4096 parallel environments at 30 Hz; all policies are deployed at 30 Hz. The benchmark contains six task categories: pick-up, tool use, peg insertion, in-hand rotation, in-hand translation, and bimanual handover.

B. Q1: Multi-Task Dexterous Manipulation in Simulation

We evaluate the state-based teacher against a ManipTrans-style imitation baseline [12] and a kinematic-retargeting baseline (Table I). Under identical settings, with 10 spatial variants per demonstration and deterministic rollouts, DEX-X reaches 65.9% average success, compared with 21.9% for ManipTrans and 5.2% for kinematic retargeting. The margin is largest on contact-rich tasks such as tool use (74.3 vs. 35.0), in-hand translation (64.8 vs. 11.5), and bimanual handover (86.6 vs. 4.7), while the two methods are closer on in-hand rotation. We attribute the gap to the absence of contact feedback in the baseline, which reproduces hand motion without maintaining stable contact forces; we frequently observe objects slipping from the hand during execution. DEX-X is instead trained in closed loop with simulated force feedback, which allows it to adjust contact throughout manipulation.

C. Q2: The Role of Scene and Tactile Representations

We ablate the scene and tactile representations of the deployable student policy, reporting average success under a strict metric (final object pose within 3 cm of the target without early termination) and a relaxed metric (5 cm). Two findings emerge. First, geometry matters: point clouds with

TABLE I: Per-category success rate (%). DEX-X stage columns report entered-stage success, while baselines report overall average. *Peg insertion uses a <1 cm criterion.

	DEX-X (ours)			ManipTrans	Kin. Retarget
Category	Grasp	Manip	Overall	Overall	Overall
Pick Up	97.4	89.9	**89.6**	1.5	1.8
Peg Insertion*	43.7	43.3	**43.1**	—	0.0
Tool Use	82.3	74.6	**74.3**	35.0	24.1
In-hand Rotation	—	36.8	**36.8**	56.9	0.0
In-hand Translation	—	64.8	**64.8**	11.5	0.0
Bimanual Handover	88.8	87.9	**86.6**	4.7	—
Average	78.0	66.2	**65.9**	21.9	5.2

local 3D force reach 44% strict success, compared with 32% for the depth input with force, indicating that an explicit 3D representation is more effective than depth images in our setting. Second, contact information matters more than its parameterization: removing force from the point-cloud policy lowers the relaxed average from 58% to 45%, with the largest reduction on contact-sensitive tasks such as pick-up and tool manipulation, yet the four encodings we tested (proprioceptive force, scalar magnitude, binary contact, and local 3D force) all improve over the no-force baseline and perform comparably to one another. We therefore adopt local 3D force vectors as the default, since they retain the most complete contact signal at comparable performance. Per-task results are reported in Appendix A.

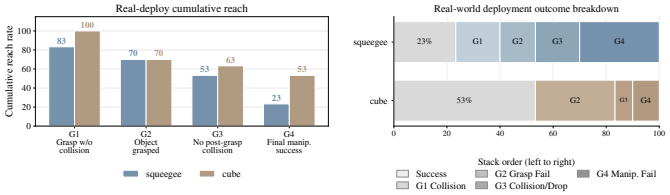

Fig. 3: **Real-world deployment** on the squeegee and cube tasks ($n = 30$ each, manually labeled). (a) Cumulative reach rate across execution stages; the cube task maintains a higher final success rate, while the squeegee task drops at the manipulation stage. (b) Outcome breakdown: squeegee failures are mainly manipulation failure and collision/drop events, whereas cube failures are dominated by grasp failures.

D. Q3: Sim-to-Real Transfer

We evaluate whether the distilled visual-tactile policy zero-shot transfers to the real platform on two representative tasks, squeegee manipulation and cube picking, recording and manually annotating execution outcomes with stage-wise failure decomposition over 30 trials each (Fig. 3). The cube task achieves 53% and the squeegee task 23%. Early-stage behaviors such as reaching and grasping transfer reliably, whereas later-stage dexterous interaction degrades due to increased sensitivity to sim-to-real mismatch and accumulated errors. Overall, the learned policies demonstrate zero-shot sim-to-real transfer to a high-DoF dexterous manipulation platform.

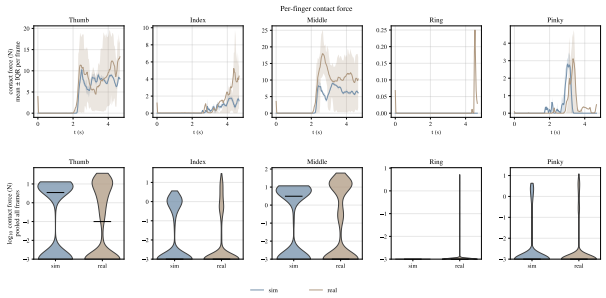

Fig. 4: **Fingertip force consistency between simulation and reality** during cube picking. Per-fingertip contact forces follow consistent contact timing, force allocation, and modulation across domains, indicating that the policy transfers contact strategies—not just motion—learned in simulation.

E. Sim-to-Real Behavior Consistency

Task success alone does not reveal whether the recovered tactile signal is what transfers. We therefore examine the contact behavior itself. For cube picking, we record per-fingertip contact forces over rollouts in both domains and compare their distributions (Fig. 4). The force profiles are consistent across simulation and reality in contact-onset timing, force allocation across fingers, and force modulation during grasp stabilization. Real measurements are noisier, which we attribute to sensing uncertainty and hardware imperfections, but the overall structure of the profiles is preserved. This suggests that the policy transfers the contact strategy acquired in simulation, and not only the kinematic motion. Motion tracking is consistent as well: real-world wrist trajectories follow the retargeted references with a mean error of 2.83 cm and a cross-episode standard deviation of 0.46–0.82 cm, indicating repeatable execution across trials. A more detailed analysis of the wrist trajectories is provided in Appendix B.

V. CONCLUSION

We introduced DEX-X, a framework that leverages simulation to recover the missing tactile supervision in human videos, enabling the learning of visual-tactile dexterous manipulation policies from monocular demonstrations. The resulting policies achieve zero-shot sim-to-real transfer of contact-rich manipulation on a 29-DoF dexterous hand-arm platform and substantially outperform imitation and kinematic-retargeting baselines. Future work includes improving deployment robustness through enhanced hardware calibration, richer tactile representations, and broader task and object coverage.

REFERENCES

[1] Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping, 2023. URL https://arxiv.org/abs/2312.02975.

[2] Rosy Chen, Mustafa Mukadam, Michael Kaess, Tingfan Wu, Francois R Hogan, Jitendra Malik, and Akash Sharma. Ptld: Sim-to-real privileged tactile latent distillation for dexterous manipulation, 2026. URL https://arxiv.org/abs/2603.04531.

[3] Yuanpei Chen, Chen Wang, Yaodong Yang, and C. Karen Liu. Object-centric dexterous manipulation from human motion data, 2024. URL https://arxiv.org/abs/2411.04005.

[4] Zerui Chen, Shizhe Chen, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Vividex: Learning vision-based dexterous manipulation from human videos, 2025. URL https://arxiv.org/abs/2404.15709.

[5] Irmak Guzey, Haozhi Qi, Julen Urain, Changhao Wang, Jessica Yin, Krishna Bodduluri, Mike Lambeta, Lerrel Pinto, Akshara Rai, Jitendra Malik, Tingfan Wu, Akash Sharma, and Homanga Bharadhwaj. Dexterity from smart lenses: Multi-fingered robot manipulation with in-the-wild human demonstrations, 2025. URL https://arxiv.org/abs/2511.16661.

[6] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makoviichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, Yashraj Narang, Jean-Francois Lafleche, Dieter Fox, and Gavriel State. Dextreme: Transfer of agile in-hand manipulation from simulation to reality, 2024. URL https://arxiv.org/abs/2210.13702.

[7] Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization, 2025. URL https://arxiv.org/abs/2507.09160.

[8] Jialei Huang, Yang Ye, Yuanqing Gong, Xuezhou Zhu, Yang Gao, and Kaifeng Zhang. Spatially anchored tactile awareness for robust dexterous manipulation, 2026. URL https://arxiv.org/abs/2510.14647.

[9] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning, 2025. URL https://arxiv.org/abs/2410.24185.

[10] Kushal Kedia, Tyler Ga Wei Lum, Jeannette Bohg, and C. Karen Liu. Simtoolreal: An object-centric policy for zero-shot dexterous tool manipulation, 2026. URL https://arxiv.org/abs/2602.16863.

[11] Chengshu Li, Mengdi Xu, Arpit Bahety, Hang Yin, Yunfan Jiang, Huang Huang, Josiah Wong, Sujay Garlanka, Cem Gokmen, Ruohan Zhang, Weiyu Liu, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Momagen: Generating demonstrations under soft and hard constraints for multi-step bimanual mobile manipulation, 2026. URL https://arxiv.org/abs/2510.18316.

[12] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning, 2025. URL https://arxiv.org/abs/2503.21860.

[13] Jason Jingzhou Liu, Yulong Li, Kenneth Shaw, Tony Tao, Ruslan Salakhutdinov, and Deepak Pathak. Factr: Force-attending curriculum training for contact-rich policy learning. *arXiv preprint arXiv:2502.17432*, 2025.

[14] Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. Dexmachina: Functional retargeting for bimanual dexterous manipulation, 2025. URL https://arxiv.org/abs/2505.24853.

[15] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

[16] Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. Spider: Scalable physics-informed dexterous retargeting, 2026. URL https://arxiv.org/abs/2511.09484.

[17] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning, 2017. URL https://arxiv.org/abs/1710.06542.

[18] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2025. URL https://arxiv.org/abs/2409.12259.

[19] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets, 2019. URL https://arxiv.org/abs/1908.09186.

[20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec

Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[21] Kenneth Shaw, Yulong Li, Jiahui Yang, Mohan Kumar Srirama, Ray Liu, Haoyu Xiong, Russell Mendonca, and Deepak Pathak. Bimanual dexterity for complex tasks. In *8th Annual Conference on Robot Learning*, 2024.

[22] Ritvik Singh, Arthur Allshire, Ankur Handa, Nathan Ratliff, and Karl Van Wyk. Dextrah-rgb: Visuomotor policies to grasp anything with dexterous hands, 2025. URL https://arxiv.org/abs/2412.01791.

[23] Jun Wang, Ying Yuan, Haichuan Che, Haozhi Qi, Yi Ma, Jitendra Malik, and Xiaolong Wang. Lessons from learning to spin "pens", 2024. URL https://arxiv.org/abs/2407.18902.

[24] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects, 2024. URL https://arxiv.org/abs/2312.08344.

[25] Tianhao Wu, Jinzhou Li, Jiyao Zhang, Mingdong Wu, and Hao Dong. Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy learning, 2025. URL https://arxiv.org/abs/2409.17549.

[26] Amber Xie, Haozhi Qi, and Dorsa Sadigh. Handelbot: Real-world piano playing via fast adaptation of dexterous robot policies, 2026. URL https://arxiv.org/abs/2603.12243.

[27] Teng Xue, Alberto Rigo, Bingjian Huang, Jiayi Shen, Zhengtong Xu, Nick Colonnese, and Amirhossein H. Memar. Tube diffusion policy: Reactive visual-tactile policy learning for contact-rich manipulation, 2026. URL https://arxiv.org/abs/2604.23609.

[28] Jessica Yin, Haozhi Qi, Jitendra Malik, James Pikul, Mark Yim, and Tess Hellebrekers. Learning in-hand translation using tactile skin with shear and normal force sensing, 2025. URL https://arxiv.org/abs/2407.07885.

[29] Jessica Yin, Haozhi Qi, Youngsun Wi, Sayantan Kundu, Mike Lambeta, William Yang, Changhao Wang, Tingfan Wu, Jitendra Malik, and Tess Hellebrekers. Osmo: Open-source tactile glove for human-to-robot skill transfer, 2025. URL https://arxiv.org/abs/2512.08920.

[30] Patrick Yin, Tyler Westenbroek, Zhengyu Zhang, Joshua Tran, Ignacio Dagnino, Eeshani Shilamkar, Numfor Mbiziwo-Tiapo, Simran Bagaria, Xinlei Liu, Galen Mullins, Andrey Kolobov, and Abhishek Gupta. Emergent dexterity via diverse resets and large-scale reinforcement learning, 2026. URL https://arxiv.org/abs/2603.15789.

[31] Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning, 2024. URL https://arxiv.org/abs/2410.02479.

[32] Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing, 2024. URL https://arxiv.org/abs/2312.01853.

[33] Shuqi Zhao, Ke Yang, Yuxin Chen, Chenran Li, Yichen Xie, Xiang Zhang, Changhao Wang, and Masayoshi Tomizuka. Dexctrl: Towards sim-to-real dexterity with adaptive controller learning, 2025. URL https://arxiv.org/abs/2505.00991.

[34] Zhe Zhao, Haoyu Dong, Zhengmao He, Yang Li, Xinyu Yi, and Zhibin Li. Closing the reality gap: Zero-shot sim-to-real deployment for dexterous force-based grasping and manipulation, 2026. URL https://arxiv.org/abs/2601.02778.

APPENDIX

A. Scene and Tactile Representation Ablations

This section reports the full ablations summarized in the main text (Q2). Throughout, the *strict* metric requires the final object pose within 3 cm of the target without early termination, the *relaxed* metric uses a 5 cm threshold, and rotation tasks use a 30° orientation threshold. Dark and light bars denote the strict and relaxed criteria, respectively.

Scene representation. We compare depth-image and point-cloud scene encodings, each with and without contact-force features (Fig. 5). Point clouds with local 3D force achieve the best average success (44% strict / 58% relaxed), improving the strict average from 32% for the depth-with-force input and confirming that explicit 3D geometry is more effective than depth images. Removing force from the point-cloud policy reduces the relaxed average from 58% to 45%, with the largest drops on contact-sensitive pick-up and tool-manipulation tasks, where force feedback stabilizes grasping and object interaction.

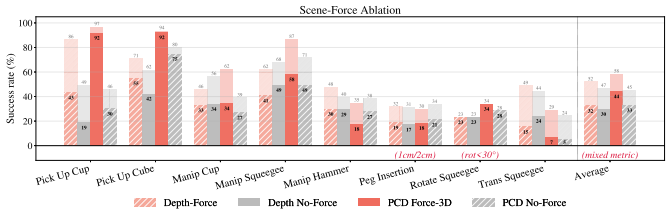

Fig. 5: **Scene-representation ablation.** Per-task success for depth vs. point-cloud scene encodings, each with and without contact force. Point clouds with local 3D force achieve the best average, and removing force hurts most on contact-sensitive tasks. Dark/light bars: strict/relaxed criteria.

Tactile representation. We compare how contact information is encoded within the point cloud: no force, proprioceptive force, scalar force magnitude, binary contact, and local 3D force vectors (Fig. 6). All tactile variants outperform the no-force baseline (36% / 47% strict/relaxed), raising the average to 43–44% / 55–56%. The different encodings perform comparably, indicating that the benefit comes from providing contact information rather than from a specific force parameterization. We adopt local 3D force vectors by default, as they preserve the most complete contact signal while remaining competitive with compact encodings.

B. Supplementary Sim-to-Real Analysis

Complementing the fingertip-force consistency reported in the main text, we further analyze spatial transfer fidelity by comparing 3D wrist trajectories between the retargeted demonstrations and real-world deployments (Fig. 7). Real trajectories closely follow the references, with a mean tracking error of 2.83 cm and a small cross-episode standard deviation (0.46–0.82 cm), indicating high repeatability. The dominant deviation is along the x-axis (the forward, load-bearing direction), while the y- and z-axes remain tightly tracked. Temporally, the

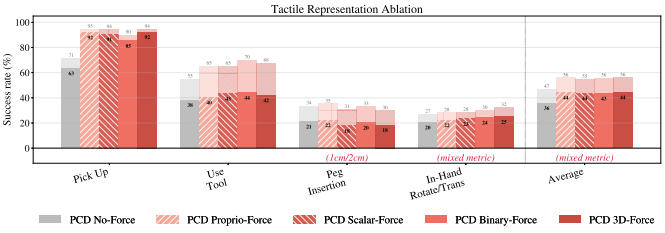

Fig. 6: **Tactile-representation ablation.** Point-cloud policies with no force, proprioceptive force, scalar magnitude, binary contact, and local 3D force. All tactile cues improve over the no-force baseline, while different encodings perform comparably. Dark/light bars: strict/relaxed criteria.

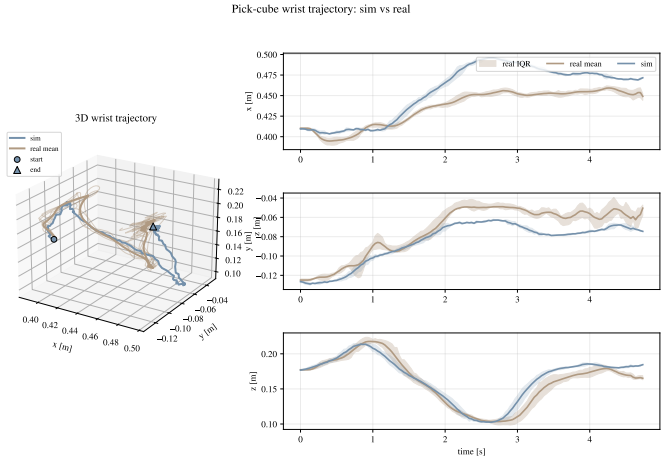

Fig. 7: **Wrist-trajectory consistency between simulation and reality** for cube picking. Real-world rollouts (mean ± IQR) closely track the retargeted reference; the dominant deviation along the load-bearing x-axis is consistent with steady-state arm-impedance compliance under payload.

error grows from 1.1 cm during approach to 2.7 cm at grasp and 3.8 cm at lift, consistent with the steady-state offset of arm impedance control under increasing payload, compounded by up to 1 cm of object-placement variation and camera-calibration error; along z the robot consistently lifts 5.7 mm higher, a uniform clearance bias. These structured patterns indicate that the residual sim-to-real gap is predominantly systematic arm compliance and perception offset rather than a learning or control failure, and could be reduced by calibrating the arm impedance gains and camera extrinsics.

C. Hand-Object Reconstruction and Retargeting

Reconstruction. Object meshes are obtained by multi-view scanning or single-image reconstruction, and per-frame 6D object poses are estimated with FoundationPose [24]. Hand keypoints are detected per view with WiLoR [18]; we fit the MANO model by minimizing multi-view 2D reprojection error over pose, root translation, and global rotation, then refine the full sequence jointly with a temporal-continuity term. A final joint hand-object optimization removes interpenetration using

a mesh-SDF penetration loss together with registration and velocity/acceleration smoothness terms, optionally constrained by recorded depth.

Retargeting. Human targets are transformed into the robot workspace and finger targets rescaled by the robot-to-human segment-length ratio (clipped to $[0.7, 1.5]$). We solve for $\mathbf{q}^{\mathrm{arm}} \in R^7$ and $\mathbf{q}^{\mathrm{hand}} \in R^{22}$ in stages: damped least-squares IK initialization on the wrist, arm-only wrist refinement, coarse finger-direction optimization (proximal joints only), and a fine all-joint stage with a Huber fingertip-tracking loss plus wrist-alignment and regularization terms. Per-finger tracking weights emphasize the thumb and index fingertips (Table II). After optimization, joint changes are capped, Savitzky–Golay smoothed, and clamped to the *measured* reachable range of the physical hand motors (tighter than the URDF limits due to tendon coupling), so training targets and deployment commands share the same feasible range. Sequences whose mean wrist tracking error exceeds a reachability threshold are discarded.

TABLE II: Per-body fingertip-tracking weights used in retargeting.

	Thumb	Index	Middle	Ring	Pinky
Fingertip	25	15	10	7	5
Proximal	3	3	3	3	3
Other	1	1	1	1	1

Spatial augmentation. To diversify initial conditions, each demonstration is augmented before retargeting with a shared planar transformation applied around the robot arm base:

$$\tilde{\mathbf{x}}_{xy} = \mathbf{R}_z(\Delta\psi)\,(\mathbf{x}_{xy} - \mathbf{c}_{xy}) + \mathbf{c}_{xy} + \Delta\mathbf{p}_{xy}, \tag{2}$$

with translation $\Delta\mathbf{p}_{xy} \sim \mathcal{U}([-5\,\mathrm{cm}, 5\,\mathrm{cm}]^2)$ and yaw $\Delta\psi \sim \mathcal{U}(-10^\circ, 10^\circ)$, where $\mathbf{c}_{xy}$ is the arm-base position. The same transform is applied to the wrist, MANO, and object trajectories, preserving relative contact geometry; augmented sequences with mean wrist IK error exceeding $3\,\mathrm{cm}$ are discarded.