Touch Improves Learning Dexterous Grasping from Demonstration

Niklas Leukroth[1*], Pankhuri Vanjani[2*], Jakub Suliga[2], Ken Nakahara[1], Rudolf Lioutikov[2,3], Roberto Calandra[1]

[1]LASR, TU Dresden [2]Intuitive Robots Lab (IRL), Karlsruhe Institute of Technology (KIT)
[3]Robotics Institute Germany (RIG)

***Abstract*—Robotic dexterous grasping with human-level performance is still one of the great unsolved challenges in robotics. Imitation Learning (IL) offers a compelling paradigm to quickly learn grasping policies by leveraging human demonstrations. Yet current approaches relying on Vision-only are brittle under occlusions, and novel objects. We present a systematic comparative study of the role of touch sensing in dexterous, multi-fingered IL, evaluating three IL algorithms (ACT, Flow Matching, BESO) across transformer, Mamba, and xLSTM backbones, and multiple tactile representations (no tactile, end-to-end CNN, and pretrained Sparsh-X encodings). Our results show that tactile signals match or improve grasping performance in 35 out of 36 scenarios over vision-only baselines, with the benefit most pronounced on unseen objects. The best configuration 'Flow Matching with an xLSTM backbone and Sparsh-X encoding' achieves the highest performance across all conditions, generalizing to a dataset of 20 novel objects from a training set of just 2 objects with a grasping success of 79%. These results provide strong evidence for the value of touch sensing towards human-level dexterous manipulation, and offer practical insights into the choice of learning algorithm, backbone, and tactile representation.**

I. Introduction

While physical interactions in our daily lives are designed specifically for human hands, the field of robotics has primarily worked in well-defined factory settings, seldom moving beyond a two-fingered gripper. With the ever-growing interest in moving robotics outside the confines of these factories and having them interact with the real-world, the need for generalization and robustness to environmental changes emerges. A growing field of research focuses on the control of anthropomorphic robotic hands, grounding robotic manipulation in the everyday interactions for which our environment is built.

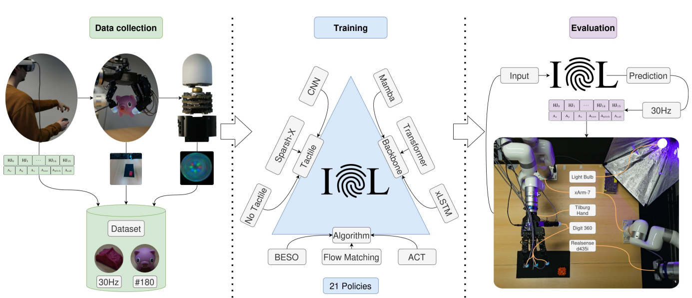

Figure 1: We study the applicability of tactile sensing for robotic dexterous grasping. After recording a multimodal dataset through teleoperation, we provide a large study over various Imitation Learning (IL) algorithms, backbones and tactile representations. Results show that tactile sensing significantly improved the generalization and robustness.

Humans rely on their sense of touch to understand the world around them, motivating research into digitizing tactile signals and incorporating them into robotic systems. Prior works have shown an increase in robustness and generalization performance on robotic grippers, suggesting this to be an interesting topic to include in robotic hands [12]. The increased complexity of dexterous hands compared to simple parallel grippers has also driven growing interest in IL algorithms that learn directly from recorded human demonstrations, which in turn necessitates good teleoperation pipelines and expensive human datasets. The designed algorithms should therefore be sample efficient, an area where touch has shown promising results [22]. Following this assessment, we investigate the influence of tactile sensing in robotic tasks on both raw performance and sample-efficient learning, including its role in generalization to novel objects.

While purely vision based robotic configurations are able to solve complex tasks in robotics, their performance often degrades under occlusion or novel objects, since visual features are tied to appearance rather than contact state [7]. Digitizing a sense of touch helps by providing appearance-invariant information in such situations of visual ambiguity [10]. Vision-based tactile sensors such as GelSight [23], Digit [9], and Digit360 [10] see wide adoption due to their spatial resolution and ability to capture rich contact geometry. While the Digit360 is a multimodal sensor, we use it here as a pure vision-based tactile sensor, making our findings applicable to other comparable vision-only sensors. Controlling anthropomorphic hands rather than grippers introduces a higher-dimensional action space, shifting the challenge toward learning manipulation behavior directly from data [13], where IL presents a compelling paradigm, directly leveraging human demonstrations [18] [17], naturally encoding human-like motion strategies and contact sequences [2] [17], and avoiding the reward-engineering challenges of Reinforcement Learning (RL) [6]. Collecting such demonstrations typically relies on teleoperation pipelines [24] [17] [15], via motion capture hardware [17] or VR tracking [15], the latter being the approach used in this work. For our study, we focus on three widely used IL algorithms: ACT [25], BESO [16], and Flow Matching [11], described further in Section III; while studied individually, no prior work systematically compares them under identical conditions on a dexterous hand.

Tactile sensing has been shown to provide contact-grounded observations that remain informative regardless of visual conditions [21], [5], but these results are primarily obtained on robotic grippers, leaving open whether the same benefits extend to dexterous multi-fingered hands with higher-dimensional action spaces [14] [19] [4] [20]. Beyond whether to include touch, how to represent tactile signals remains an open design question [5]: end-to-end approaches feed raw tactile images directly into the policy network, while pretrained embeddings like Sparsh-X compress tactile signals into representations learned from large tactile datasets, offering potentially better generalization [5]. In contrast to prior work that studies tactile sensing in isolation within single-algorithm, gripper-based settings, we provide a systematic comparison across multiple IL algorithms, backbone architectures, and tactile representations on a dexterous multi-fingered hand, allowing the contribution of touch to be evaluated independently of architectural choices.

By evaluating multiple state-of-the-art IL algorithms across transformer, state-space-model, and recurrent backbones, alongside a range of tactile representations, we shift the focus onto the tactile sensor itself rather than any single method.

The key contributions of this work include:

1) Evaluation of grasping accuracy across state-of-the-art IL algorithms, backbones and tactile representations
2) We perform an extensive evaluation on a dataset of 20 novel objects, testing generalization capabilities of policies that were trained on a two object dataset

II. Experimental Setup

Hardware Platform and Data Collection Pipeline

To collect human demonstrations, we developed a VR based teleoperation system using the Meta Quest 3's internal hand and finger tracking, placing virtual trackers on the fingers and wrist of the operator. This allows us to collect a large dataset of real world demonstrations in a time-efficient manner while featuring full dexterity. Collecting data this way has the benefit of only containing state-action pairs that are feasible, while incorporating accurate sensor modalities. Fig. 1 shows the hardware setup used during data collection and evaluation. For the robotic hardware, we used an xArm-7 robot arm mounting a Tilburg hand equipped with four Digit360 [10] tactile sensors at the finger tips. Completing the setup is a camera that provides a side-angle view.

Dataset The *Known* dataset consists of two objects, shown in Fig. 2 that are made of rather firm plastic and are comparatively larger in size than the novel objects. Both objects are being grasped ten times in nine different positions, for a total of 180 grasps. We recorded the dataset at a time-synchronized rate of 30 Hz, while cropping and downscaling both the camera and tactile images to a resolution of $224 \times 224 \times 3$. The action

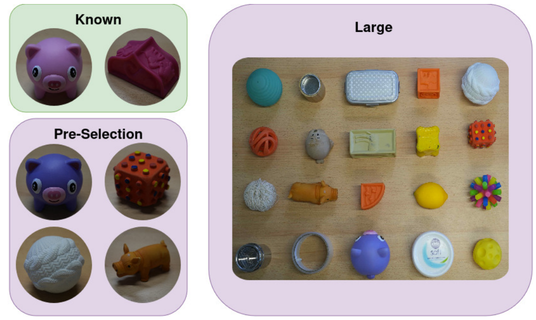

Figure 2: The dataset is split into a *Known* set, which the policies are trained on, a *Pre-selection* set that every policy is evaluated on and a *Large* dataset, against which the best performing algorithms are tested.

space is a $\mathbb{R}^{22}$ vector, that consists of a $\mathbb{R}^{16}$ vector representing the 16 joint angles of the hand and a $\mathbb{R}^{6}$ vector for the end-effector position of the xArm-7. Given this data, we train 21 IL policies with different combinations of backbones and tactile representations, evaluating them over different dataset sizes. We attribute the binary classifier for success to a grasping attempt if the policy is able to pick up the object and hold it in the air for three seconds within a total time limit of 13 seconds.

III. Policy Design and Tactile Integration

Imitation Learning Methods We benchmark our approach against a suite of imitation learning policies drawn from X-IL [8]. X-IL framework decouples backbone architecture from action representation, allowing controlled comparisons across configurations. We evaluate three backbone architectures: a **Transformer**, which uses multi-head self-attention to model temporal dependencies across observation histories, **Mamba** [3], which replaces attention with selective state-space models that compress history into a fixed-size recurrent state with linear complexity, and **xLSTM** [1], which extends classical LSTMs with exponential gating and matrix memory cells for improved long-range memory retention. For the action generation head, we pair each backbone with **Flow Matching**, which maps Gaussian noise to actions along straight interpolation paths, and **BESO**, a score-matching diffusion method that iteratively denoises actions toward the target distribution. Additionally, we evaluate **ACT** (Action Chunking with Transformers) [25], which predicts chunks of future actions at each timestep using a CVAE-based transformer, mitigating compounding errors, and using temporal ensembling to smooth the executed actions by averaging overlapping predicted chunks.

Tactile sensing integration through learning We focus on the Digit360 sensor, that provides sensitive high resolution RGB images. To better understand the influence of different tactile sensing strategies, we evaluate three approaches. The **Vision** baseline omits tactile observations entirely during both training and inference, serving as a reference to quantify the overall contribution of touch sensing. In the **Vision+Tactile (CNN)** approach, the raw RGB tactile images are encoded by a lightweight convolutional network trained from scratch jointly with the policy, requiring no prior tactile data. Finally, we use **Vision+Tactile (Sparsh-X)** [5], a pretrained transformer-based encoder specifically trained on Digit360 data, which produces compact tactile embeddings without requiring task-specific encoder training. This allows us to disentangle the effect of tactile sensing itself from the choice of representation, and to assess whether pretrained tactile encoders offer an advantage over task-specific learned ones.

IV. Experimental Evaluation

To systematically study the influence of tactile sensing, we train each policy variant under identical conditions, varying only the tactile representation while holding all other parameters fixed. Evaluation is conducted under the same physical setup as data collection, including the initial starting pose of the robotic hand, ensuring that observed differences reflect policy behavior rather than environmental variation. In order to compare the general grasping performance of the different policies, they are presented with the six objects that make up the combination of the *Known* and *Pre-Selection* dataset, seen in Fig. 2. This controlled setup allows us to isolate the contribution of touch across the following research questions:

A. Performance and Generalization: How does tactile sensing affect grasping performance and generalization to novel objects in dexterous imitation learning?

To evaluate the grasping performance achieved for every possible combination of IL algorithm, backbone and tactile representation, every trained policy grasps an object nine times in different positions, totaling 54 grasps per policy. Fig. 3 differentiates the results achieved over the different learning algorithms, while separating backbone architecture, representation and performance split between known and novel objects. With one exception for Flow Matching with a transformer backbone that displays a drop in performance using the Sparsh-X encoding on known objects compared to the no tactile signal option, tactile signals match or outperform the no-tactile baseline across 27 out of 28 comparisons in this experiment. This benefit is most pronounced on novel objects, where tactile variants always improve over the vision-only baseline, suggesting that contact information supports generalization beyond the training distribution. The weakest overall performance is observed with BESO using a transformer backbone, where both tactile variants reach only 56% success rate. Notably, the corresponding no-tactile policy fails to complete a single successful grasp in this configuration, on either known or novel objects.

Evaluation on *Large* dataset On the *Pre-Selection* dataset, the best non-tactile and tactile policies for novel objects were both Flow Matching with an xLSTM backbone. We therefore took these two policies, trained on 180 grasps from the *Known* dataset, and evaluated them on the bigger *Large* dataset (Fig. 2). This dataset has 20 novel objects, each grasped nine times, for 180 attempts per policy. Fig. 4 shows the resulting performance and its distribution over object weight and size. The vision-only policy reaches 53% success, compared to 79% with Sparsh-X tactile encoding.

Since both policies share the same algorithm and backbone, this is a direct comparison of the effect of tactile sensing. Outside the highlighted region marking the *Known* training objects' characteristics, the non-tactile policy's performance drops for larger, heavier objects, while the tactile policy stays evenly distributed. This matches expectations: larger, heavier objects need more accurate contact-patch placement, which touch provides, while smaller objects can be enclosed by the hand.

B. How does the choice of learning algorithm and backbone architecture affect tactile imitation learning performance?

The performance of ACT, Flow Matching, and BESO is shown in Fig. 3. Although ACT and BESO each perform well in their best configurations, the highest grasping performance on both known and novel objects is achieved by Flow Matching with an xLSTM backbone and Sparsh-X tactile encoding, which we attribute to Flow Matching's ability to learn smooth, straight trajectories well-suited to precise, continuous grasping, whereas BESO's iterative denoising introduces stochasticity less suited to fine motor consistency. Since BESO and Flow Matching were each trained with the Transformer, Mamba, and xLSTM backbones, we analyze their performance differences further in Fig. 5. The Transformer consistently underperforms both Mamba and xLSTM, as without explicit state compression it must learn entirely from attention which timesteps carry contact-relevant information, requiring more data to converge. Mamba and xLSTM instead compress the observation history into a hidden state, giving a more natural inductive bias for retaining task-relevant signals and converging reliably from limited demonstrations. Between Mamba and xLSTM the difference is more nuanced: under BESO the two perform similarly, but xLSTM outperforms Mamba with Flow Matching, suggesting its structured gating complements Flow Matching's straight-path action generation, making this the best-performing configuration overall.

V. Conclusion

In this work, we study the role of touch sensing for robotic dexterous grasping through IL, examining performance, generalization and robustness to novel objects. Flow Matching [11] with an xLSTM backbone and Sparsh-X tactile encoding consistently performed best, and Mamba [3] and xLSTM [1] backbones outperformed the transformer baseline across both algorithms. Touch sensing improved grasping success in 35 of 36 test cases and substantially increased robustness. It also enabled a policy trained on just 180 demonstrations from 2 objects to generalize to 20 novel objects with 79% success.

These results suggest touch provides a representational space well suited to grasping, which can be viewed as balancing contact forces against gravity. Touch yields direct, low-noise contact-force information, whereas vision provides only weak indirect cues, and is invariant to appearance, background,

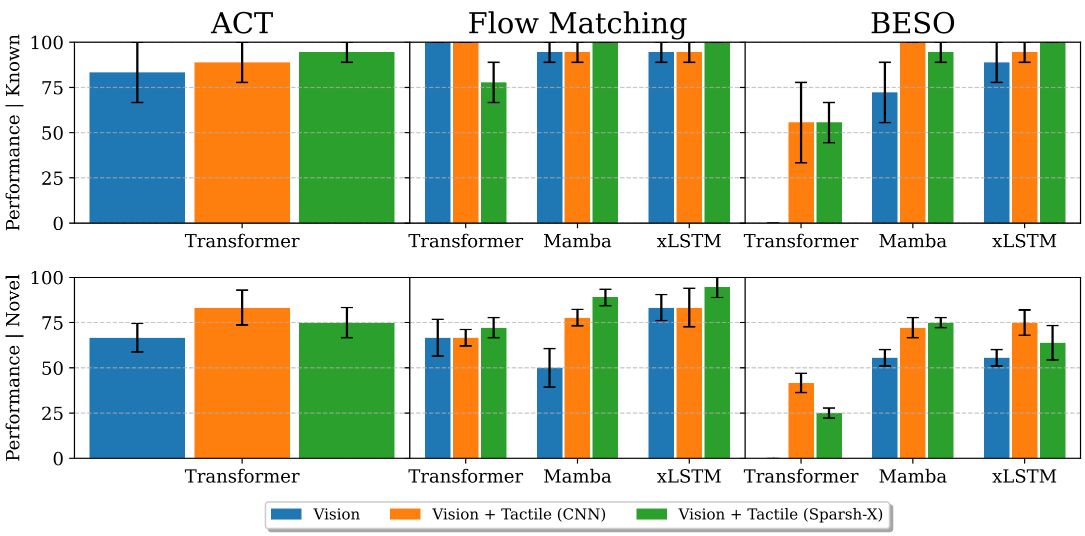

Figure 3: Grasping performance on known and novel objects with Standard Error of the Mean (SEM), compared across policies and backbones. Including tactile sensing improves grasping performance, especially on novel objects. (BESO + Transformer + Vision-only: 0% success, bar absent.)

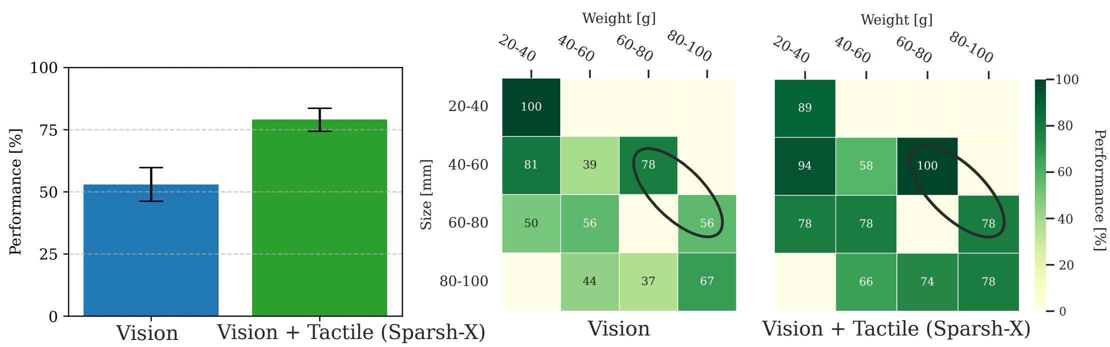

Figure 4: Grasping performance on the *Large* dataset (20 novel objects, SEM), and performance distribution over object weight and size. Vision + Tactile (Sparsh-X) outperforms Vision only by 25% over 180 grasping attempts, with the gap widening for larger, heavier objects. Circled areas mark the training objects' characteristics.

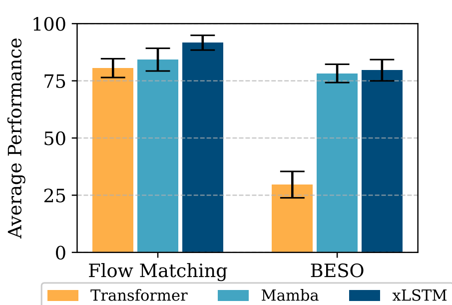

Figure 5: Average performance of different backbones $\in$ {Transformer, Mamba, xLSTM} for BESO and Flow Matching with SEM. Mamba and xLSTM consistently outperform Transformer, with xLSTM slightly ahead of Mamba, though the difference is too marginal for conclusive assessments.

and occlusion factors that otherwise hinder vision-only policies. Our study has three main limitations that motivate future work. First, the demonstrated grasps were limited to top-down, fingertip only motions, whereas everyday human grasping involves richer trajectories (e.g., side grasps) and grasp types. Second, while the *Large* evaluation dataset includes some household items, most objects are brightly colored and toy-like, motivating further evaluation on more diverse, realistic objects. Third, our evaluation focuses on grasping; extending this analysis to tasks demanding higher manipulation precision, such as in-hand rotation, would further clarify the benefits of tactile sensing for general dexterous manipulation.

References

[1] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ARAxPPIAhq.

[2] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12248–12255, 2025. doi: 10.1109/IROS60139.2025.11247017.

[3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

[4] Erik Helmut, Niklas Funk, Tim Schneider, Cristiana de Farias, and Jan Peters. Tactile-conditioned diffusion policy for force-aware robotic manipulation, 2025. URL https://arxiv.org/abs/2510.13324.

[5] Carolina Higuera, Akash Sharma, Taosha Fan, Chaithanya Krishna Bodduluri, Byron Boots, Michael Kaess, Mike Lambeta, Tingfan Wu, Mustafa Mukadam, Zixi Liu, and Francois Robert Hogan. Tactile beyond pixels: Multisensory touch representations for robot manipulation. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=sMs4pJYhWi.

[6] Yu Ishihara, Noriaki Takasugi, Kotaro Kawakami, Masaya Kinoshita, and Kazumi Aoyama. Constraints as rewards: Reinforcement learning for robots without reward functions, 2025. URL https://arxiv.org/abs/2501.04228.

[7] Jaehwan Jang, Byeong-Sun Park, Kyeong Taek Oh, Seong-Jae Yoo, Seong-Min Im, Yasser Khan, and Min-gu Kim. Complementary visual localization and tactile mapping approach for robotic perception of millimeter-sized objects with irregular surfaces. *Microsystems & Nanoengineering*, 12(1):91, Mar 2026. ISSN 2055-7434. doi: 10.1038/s41378-026-01190-8. URL https://doi.org/10.1038/s41378-026-01190-8.

[8] Xiaogang Jia, Atalay Donat, Xi Huang, Xuan Zhao, Denis Blessing, Hongyi Zhou, Han A Wang, Hanyi Zhang, Qian Wang, Rudolf Lioutikov, et al. X-il: Exploring the design space of imitation learning policies. *arXiv preprint arXiv:2502.12330*, 2025.

[9] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. ISSN 2377-3774. doi: 10.1109/lra.2020.2977257. URL http://dx.doi.org/10.1109/LRA.2020.2977257.

[10] Mike Lambeta, Tingfan Wu, Ali Sengul, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, Norb Tydingco, Gregg Kammerer, Dave Stroud, Jake Khatha, Kurt Jenkins, Kyle Most, Neal Stein, Ricardo Chavira, Thomas Craven-Bartle, Eric Sanchez, Yitian Ding, Jitendra Malik, and Roberto Calandra. Digitizing touch with an artificial multimodal fingertip, 2024. URL https://arxiv.org/abs/2411.02479.

[11] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[12] Fangchen Liu, Chuanyu Li, Yihua Qin, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface, 2025. URL https://arxiv.org/abs/2504.06156.

[13] Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibei Yang, Wenping Wang, Jingyi Yu, Xuming He, and Yuexin Ma. Realdex: towards human-like grasping for robotic dexterous hand. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/758. URL https://doi.org/10.24963/ijcai.2024/758.

[14] Quan Khanh Luu, Pokuang Zhou, Zhengtong Xu, Zhiyuan Zhang, Qiang Qiu, and Yu She. Manifeel: Benchmarking and understanding visuotactile manipulation policy learning, 2026. URL https://arxiv.org/abs/2505.18472.

[15] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.015.

[16] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-Conditioned Imitation Learning using Score-based Diffusion Policies. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.028.

[17] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS)*, 2024. URL https://openreview.net/forum?id=EiarCRjOm7.

[18] Edgar Welte and Rania Rayyes. Interactive imitation learning for dexterous robotic manipulation: challenges and perspectives—a survey. *Frontiers in Robotics and AI*, 12, December 2025. ISSN 2296-9144. doi: 10.3389/frobt.2025.1682437. URL http://dx.doi.org/10.3389/frobt.2025.1682437.

[19] Yansong Wu, Zongxie Chen, Fan Wu, Lingyun Chen, Liding Zhang, Zhenshan Bing, Abdalla Swikir, Sami Haddadin, and Alois Knoll. Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11831–11837. IEEE, 2025.

[20] Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang,

Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[21] Qi Ye, Qingtao Liu, Siyun Wang, Jiaying Chen, Yu Cui, Ke Jin, Huajin Chen, Xuan Cai, Gaofeng Li, and Jiming Chen. Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity. *Science Robotics*, 11(110):eady2869, 2026. doi: 10.1126/scirobotics.ady2869. URL https://www.science.org/doi/abs/10.1126/scirobotics.ady2869.

[22] Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=7yMZAUkXa4.

[23] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12), 2017. ISSN 1424-8220. doi: 10.3390/s17122762. URL https://www.mdpi.com/1424-8220/17/12/2762.

[24] Maryam Zare, Parham M. Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 54(12):7173–7186, Dec 2024. ISSN 2168-2275. doi: 10.1109/TCYB.2024.3395626.

[25] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016.