KineDex: Learning Tactile-Informed Visuomotor Policies via Kinesthetic Teaching for Dexterous Manipulation

Di Zhang[1,5] * Chengbo Yuan[2,5,6] Chuan Wen[3] Hai Zhang[4,5]
Junqiao Zhao[1] Yang Gao[2,5,6]†

[1]Tongji University [2]Tsinghua University [3]Shanghai Jiao Tong University
[4]University of Hong Kong [5]Shanghai Qi Zhi Institute [6]Shanghai AI Lab

https://dinomini00.github.io/KineDex/

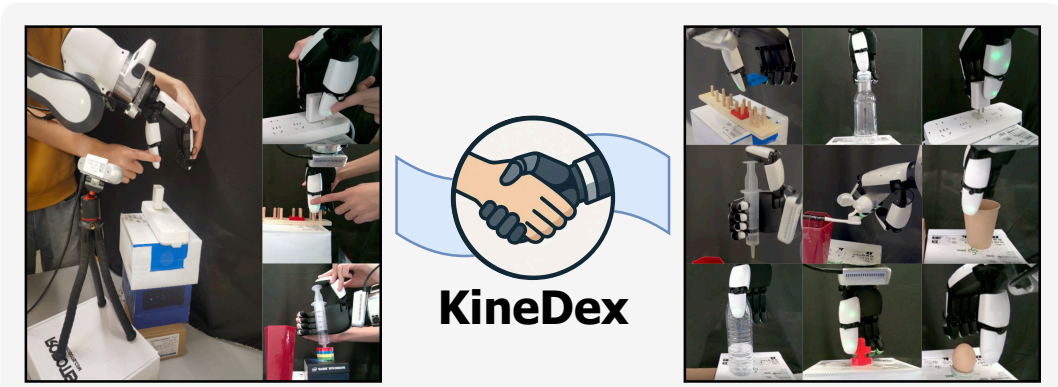


Figure 1: We present **KineDex**, a framework for collecting tactile-enriched demonstrations via kinesthetic teaching and training tactile-informed visuomotor policies for dexterous manipulation.

Abstract: Collecting demonstrations enriched with fine-grained tactile information is critical for dexterous manipulation, particularly in contact-rich tasks that require precise force control and physical interaction. While prior works primarily focus on teleoperation or video-based retargeting, they often suffer from kinematic mismatches and the absence of real-time tactile feedback, hindering the acquisition of high-fidelity tactile data. To mitigate this issue, we propose **KineDex**, a hand-over-hand kinesthetic teaching paradigm in which the operator's motion is directly transferred to the dexterous hand, enabling the collection of physically grounded demonstrations enriched with accurate tactile feedback. To resolve occlusions from human hand, we apply inpainting technique to preprocess the visual observations. Based on these demonstrations, we then train a visuomotor policy using tactile-augmented inputs and implement force control during deployment for precise contact-rich manipulation. We evaluate KineDex on a suite of challenging contact-rich manipulation tasks, including particularly difficult scenarios such as squeezing toothpaste onto a toothbrush, which require precise multi-finger coordination and stable force regulation. Across these tasks, KineDex achieves an average success rate of 74.4%, representing a 57.7% improvement over the variant without force control. Comparative experiments with teleoperation and user studies further validate the advantages of KineDex in data collection efficiency and

*Email: 2331922@tongji.edu.cn
†Corresponding at: gaoyangiiis@mail.tsinghua.edu.cn

operability. Specifically, KineDex collects data over twice as fast as teleoperation across two tasks of varying difficulty, while maintaining a near-100% success rate, compared to under 50% for teleoperation.

Keywords: Dexterous Manipulation, Kinesthetic Teaching, Tactile Sensing

1 Introduction

Integrating tactile sensing with dexterous hands substantially improves robotic manipulation capabilities [1, 2, 3, 4], paving the way for broader deployment in daily scenarios. Despite notable advances in recent years, particularly in hardware offering increased flexibility [5, 6, 7, 8] and improved tactile sensor precision [9, 10, 11, 12], acquiring expert-level demonstrations that incorporate high-fidelity tactile sensing remains a fundamental challenge.

Most existing demonstration collection methods focus on human hand motion retargeting. A common approach is teleoperation [13, 14, 15, 16, 17, 18], which captures the demonstrator's hand trajectories using a virtual reality headset or data gloves and maps them to robotic hands. Another line of work directly leverages egocentric videos to infer corresponding robot motions [19, 20, 21, 22, 23]. However, these methods face limitations when applied to complex manipulation tasks. First, the retargeted trajectories often fail to accurately reproduce human behavior due to the kinematic mismatch between human and robotic hands. Second, the absence of on-robot tactile feedback during teleoperation makes task performance highly dependent on the operator's expertise. To alleviate this issue, recent work has introduced exoskeleton-based systems that provide real-time haptic feedback as a proxy for touch [24, 25]. However, the interaction experience still differs notably from direct physical contact, and the data collection efficiency remains comparable to that of traditional teleoperation.

Motivated by recent advances in the biomimetic design of dexterous hands [26], we propose **KineDex**, a framework for collecting tactile-enriched demonstrations through hand-over-hand guidance. This paradigm offers several key advantages: (i) human motions can be directly applied to the robotic hand, eliminating retargeting errors; (ii) operators receive precise force feedback, enabling the collection of high-quality tactile data; (iii) data collection efficiency approaches that of direct human execution; and (iv) the framework naturally extends to more complex hardware and challenging contact-rich dexterous manipulation tasks, addressing limitations of previous kinesthetic teaching setups [27, 28, 29].

Building upon the tactile-enriched kinesthetic demonstrations, we further leverage them for visuomotor policy training. However, this remains challenging due to the domain shift introduced by the presence of the human hand in collected visual observations during training, but absent at inference time. To address this, we apply inpainting techniques [30] to remove occlusions, providing a more scalable and efficient alternative to prior trajectory replay-based methods [28]. To better perform contact-rich manipulation, we incorporate tactile sensing to enrich the policy inputs. Moreover, the policy is trained to predict target fingertip forces alongside target joint positions, which we refer to as force-informed actions, enabling KineDex to achieve accurate force control during execution.

We conduct extensive experiments on nine contact-rich manipulation tasks, as illustrated in Figure 1. KineDex demonstrates robust and precise control, achieving an average success rate of 74.4% across all tasks and outperforming the variant without force control by 57.7%. In addition, we find that tactile sensing is particularly important for contact-intensive tasks such as *Cap Twisting*, *Toothpaste Squeezing*, and *Syringe Pressing*. Experimental results show that KineDex outperforms the variant without tactile sensing by 26.7% on these tasks. We conduct comparative experiments with teleoperation to further assess the efficiency of KineDex. The results show that KineDex achieves more than twice the data collection speed across two tasks of varying difficulty, while maintaining a success rate close to 100%. In contrast, teleoperation yields an average success rate below 50% and requires significantly more time to collect the same amount of data. User study feedback addi-

tionally confirms the advantages of KineDex over teleoperation, with participants reporting a more intuitive and efficient experience when using kinesthetic teaching.

In summary, we present **KineDex**, a framework for dexterous manipulation that integrates kinesthetic data collection with tactile-informed visuomotor policy training. Through empirical evaluations and user studies, we demonstrate that kinesthetic teaching provides a more effective and efficient alternative to teleoperation for collecting high-quality demonstrations with dexterous hands. Furthermore, by leveraging tactile-augmented demonstrations, we train a tactile-informed policy that incorporates force control during inference, enabling successful execution across nine distinct contact-rich manipulation tasks.

2 Related Work

2.1 Collecting Demonstrations with Dexterous Hands

Most existing methods for collecting demonstrations with dexterous hands rely on retargeting human hand motion via teleoperation [31, 32, 33, 34, 35] or using video data [19, 20, 21, 22, 23]. However, both approaches share a key limitation: the operator lacks real-time tactile feedback, adversely impacting the efficiency and success rate of data collection. To mitigate this, some studies have introduced exoskeleton systems [36, 24, 25] that simulate haptic feedback during interaction. Nevertheless, such approximations differ fundamentally from the actual tactile sensations experienced through direct physical contact.

Kinesthetic teaching [37, 29, 38, 39, 28] enables operators to physically manipulate the robotic hardware and receive real-time force feedback. HIRO [39] introduces a hand-over-hand teaching paradigm, where the demonstrator grasps the robotic hand to experience accurate force feedback, but does not record tactile data, limiting its applicability to complex tasks. The most closely related work to ours is DexForce [28], which collects tactile-augmented demonstrations via kinesthetic teaching but relies on simpler hardware with fewer degrees of freedom and evaluates only on basic tasks such as grasping or flipping. Moreover, due to visual occlusions from the demonstrator's hand, DexForce requires replaying kinesthetic trajectories to obtain clean observations—a process that becomes infeasible for longer-horizon tasks.

2.2 Learning Dexterous Manipulation from Human Demonstrations

Learning from expert human demonstrations [40, 41, 42, 43, 44, 45] enables embodied agents to acquire autonomous manipulation skills. Diffusion Policy [44], which applies diffusion models to imitation learning, has shown strong capability in capturing the multimodal structure of demonstration data. Extensions such as 3D Diffusion Policy [45] integrate point cloud information into the perception pipeline, allowing for richer environmental representations. Other works have explored the integration of tactile sensing [46, 47, 48], enabling policies to perform more fine-grained manipulation. In our work, we adopt Diffusion Policy as the backbone for policy learning, augmenting its observation space with tactile sensing and applying force control during inference to ensure precision and stability in contact-rich manipulation tasks.

3 Method

3.1 Problem Formulation

In this work, we aim to train tactile-informed visuomotor policies from kinesthetic demonstrations via imitation learning, and to implement force control at inference time for executing contact-rich manipulation tasks. The observation space and learning targets are defined as follows.

Observation Space. At each time step t, the policy receives multi-modal observations comprising: (i) RGB images captured from N_o multi-view cameras, denoted as $o_t \in \mathbb{R}^{N_o \times H \times W \times 3}$; (ii) tactile

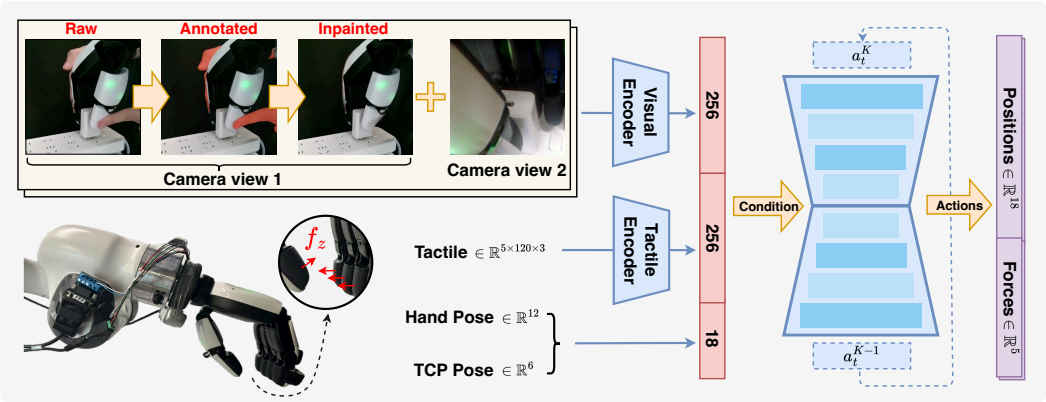


Figure 2: Overview of the **KineDex** framework. KineDex collects tactile-enriched demonstrations via kinesthetic teaching, where visual occlusions from the operator's hand are removed through inpainting before policy training. The learned policy takes visual and tactile inputs to predict joint positions and contact forces, which are executed with force control for robust manipulation.

vectors obtained from N_q sensing points on each of the five fingertips, represented as $q_t \in \mathbb{R}^{5\times N_q}$; (iii) proprioceptive signals from N_x joints of the robotic hardware, given by $x_t \in \mathbb{R}^{N_x}$.

Action Space. To enable precise force control during inference, the policy predicts not only the target joint positions $x_d \in \mathbb{R}^{N_x}$, but also the N_f-dimensional target fingertip forces $f_d \in \mathbb{R}^{5\times N_f}$, resulting in a force-informed action denoted by $a_t = \{x_d, f_d\}$. While x_d specifies the baseline configuration of the hand, f_d modulates the contact forces via a modified PD controller (introduced in Section 3.4), allowing the fingers to actively track the desired contact forces during execution.

3.2 Kinesthetic Data Collection

The general hardware setup of **KineDex** consists of a robotic arm equipped with a dexterous hand, as illustrated in Figure 1. In addition, we employ two RGB cameras to capture visual observations: one is mounted in front of the workspace to provide a global view of the scene, and the other is wrist-mounted on the end-effector to enable close-range perception of the manipulation area.

The core idea of KineDex data collection system is to allow operators to "wear" the dexterous hand while moving freely to perform precise contact-rich manipulation in real time. To enable this hand-over-hand control, we attach ring-shaped straps to the dorsal sides of the four non-thumb fingers of the robotic hand, allowing the operator to guide the hand as if wearing a glove. This physical coupling ensures that contact forces experienced during motion are immediately transmitted to the operator's hand, providing natural haptic feedback throughout the demonstration. Due to morphological differences between human and robotic hands, the operator controls the thumb separately with their left hand, while guiding the remaining fingers with their right hand as described above. Examples of kinesthetic demonstrations are shown in Figure 4.

During kinesthetic teaching, the following data modalities are recorded for each demonstration:

- *Visual observations*: RGB images captured from the front-facing and wrist-mounted cameras. As raw observations may contain the operator's body, we apply an inpainting strategy to address this issue, as detailed in Section 3.3.
- *Proprioception*: The robot arm's end-effector pose and the dexterous hand's joint positions.
- *Tactile sensing*: Per-finger tactile measurements, with each finger equipped with multiple sensing points that record localized contact forces, forming a dense tactile sensing matrix.
- *Fingertip force*: A 3D force vector $\mathbf{f} = (f_x, f_y, f_z)$ for each fingertip, where each component represents the force along the corresponding axis, computed by aggregating the localized forces from all tactile sensing points.

3.3 Policy Learning

The raw kinesthetic demonstrations collected by the system cannot be directly used for visuomotor policy learning, as the front-facing camera inevitably captures the operator's body during interaction. Training on such data introduces a significant out-of-distribution(OOD) shift at inference time, when the human body is no longer present in the scene (as shown in Table 1). Motivated by recent advances in human-to-robot data editing [49, 50], we adopt an inpainting-based approach to remove the operator's body from the visual observations.

For raw kinesthetic demonstrations, we first apply Grounded-SAM [51] to extract masks of the operator's body parts from the video frames. These frames, along with their corresponding masks, are then passed to the ProPainter [30] model to inpaint the occluded human body regions. An example of the data preprocessing pipeline is illustrated in Figure 2. Although the inpainting model is not pretrained on robot-specific data and may not achieve perfect removal, our experiments demonstrate that the resulting demonstrations are sufficient for training high-performance policies.

Using the preprocessed demonstrations, we train Diffusion Policy [44] conditioned on inpainted visual observations, tactile sensing, and proprioception to predict force-informed actions, modeled as $p(x_d, f_d \mid o_t, q_t, x_t)$. Specifically, for the calculated fingertip force vector $\mathbf{f} = (f_x, f_y, f_z)$, we supervise policy training using the normal force component f_z, which corresponds to the primary axis along which the fingertip can actively exert force, as illustrated in Figure 2.

At inference time, the policy produces action chunks [52] to enable smoother control. Each chunk specifies the desired joint positions and fingertip contact forces, which are executed using the force control strategy described below. Further details of the network architecture and policy training configurations are provided in Appendix B.

3.4 Force Control

In conventional setups, robots are typically operated under position control [53], where the control signal u is computed by a PD controller [54] based on the joint position and velocity errors relative to the target joint positions x_d:

$$u = K_p(x_d - x) + K_d(\dot{x_d} - \dot{x}), \tag{1}$$

where K_p and K_d denote the proportional and derivative gains, respectively.

However, relying solely on position control may be insufficient for certain precise, contact-rich tasks, such as *cap twisting* or *toothpaste squeezing*. This is due to policies that only track target joint positions results in merely contacting the object's surface without applying meaningful forces, as the recorded fingertip positions remain unchanged regardless of the forces applied during kinesthetic teaching. This discrepancy often leads to unstable grasps, slipping, or ineffective manipulation.

To address this limitation, we exploit the following physical property: when a fingertip contacts an object, any nonzero position error continuously generates pressure against the surface due to the object's resistance, with larger errors producing greater forces, as if the target position lies inside the object. We refer to this virtual displacement as **the force-informed target position**. To compute it, we utilize the predicted fingertip forces f_d from the trained policy, which are directed orthogonal to the contact surface and align with the primary motion axis of the finger joints. Specifically, for each finger, let x^{tip} and x^{base} denote the current positions of the fingertip and base joints, respectively. The force-informed target positions, x_d^{tip} and x_d^{base}, are then computed as:

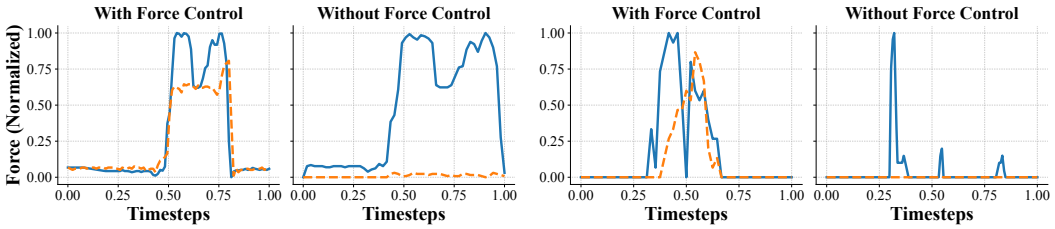
$$\begin{cases} x_d^{tip} = x^{tip} + K^{tip} \cdot f_d \\ x_d^{base} = x^{base} + K^{base} \cdot f_d \end{cases} \tag{2}$$

Here, K^{tip} and K^{base} are two hyperparameters that determine the motion stiffness at the fingertip and base joints. They are tuned to ensure that the execution faithfully tracks the predicted forces and are kept fixed across different tasks. With this force control strategy, KineDex can precisely track the target fingertip forces predicted by the trained policy, thereby achieving stable and force-informed control during execution.

Table 1: Number of successful trials (out of 20) **during inference** for different methods.

Method	*Bottle Picking*	*Cup Picking*	*Egg Picking*	*Cap Twisting*	*Nut Tightening*
KineDex	17	20	17	15	16
w/o Force Control	0	16	5	2	7
w/o Tactile Input	15	17	18	10	12
w/o Inpainting	0	0	0	0	0

Method	*Peg Insertion*	*Charger Plugging*	*Toothpaste Squeezing*	*Syringe Pressing*
KineDex	15	12	9	13
w/o Force Control	0	0	0	0
w/o Tactile Input	16	10	3	8
w/o Inpainting	0	0	0	0



(a) *Charger Plugging*. (b) *Egg Picking*.

Figure 3: Visualization of predicted and sensed forces at the thumb during task execution, comparing the force-informed policy and the variant without force control.

4 Experiments

In this section, we investigate the effectiveness of kinesthetic demonstrations for training visuomotor policies across a range of contact-rich dexterous manipulation tasks. We further evaluate the efficiency and practicality of kinesthetic teaching through comparative experiments and a user study, highlighting its advantages over teleoperation-based approaches.

To this end, we design a suite of nine tasks that emphasize precise force control, multi-finger coordination, and interaction with everyday objects. These tasks span a range of dexterous skills, including challenging scenarios such as squeezing toothpaste onto a toothbrush, which requires continuous and fine-grained pressure modulation; and pressing a syringe, which demands stable unimanual actuation and a coordinated grip to prevent slippage or misalignment. Detailed task descriptions are provided in Appendix A.

4.1 Performance Evaluation

Hardware Setup. For this set of experiments, we implement KineDex using a Franka Emika Panda robotic arm equipped with a Robotera XHand1[3] dexterous hand. Each finger on the XHand1 has two joints, while the thumb and index finger include an additional rotational joint, resulting in a total of 12 degrees of freedom. Each finger is equipped with 120 tactile sensing points.

Baselines. We compare KineDex against three ablated variants: (i) *w/o Force Control*, which disables force control during inference while keeping the training setup unchanged. (ii) *w/o Tactile Input*, which removes tactile sensing from the policy inputs during training; however, the policy still predicts target fingertip forces, which are executed using the same force control strategy. (iii) *w/o Inpainting*, which omits the inpainting preprocessing step for kinesthetic demonstrations.

We evaluate performance by conducting 20 trials per task, with results summarized in Table 1. KineDex achieves over 70% success rates on most tasks, and nearly 100% on common pick-and-place scenarios such as *Bottle Picking* and *Cup Picking*. While performance is slightly lower on the

[3] https://www.robotera.com/goods/2.html

Table 2: Number of successful trials (out of 20) **during data collection** for different methods.

Method	*Bottle Picking*	*Cup Picking*	*Cap Twisting*	*Charger Plugging*	*Syringe Pressing*
KineDex	20	20	20	18	20
Teleoperation	19	9	7	0	4

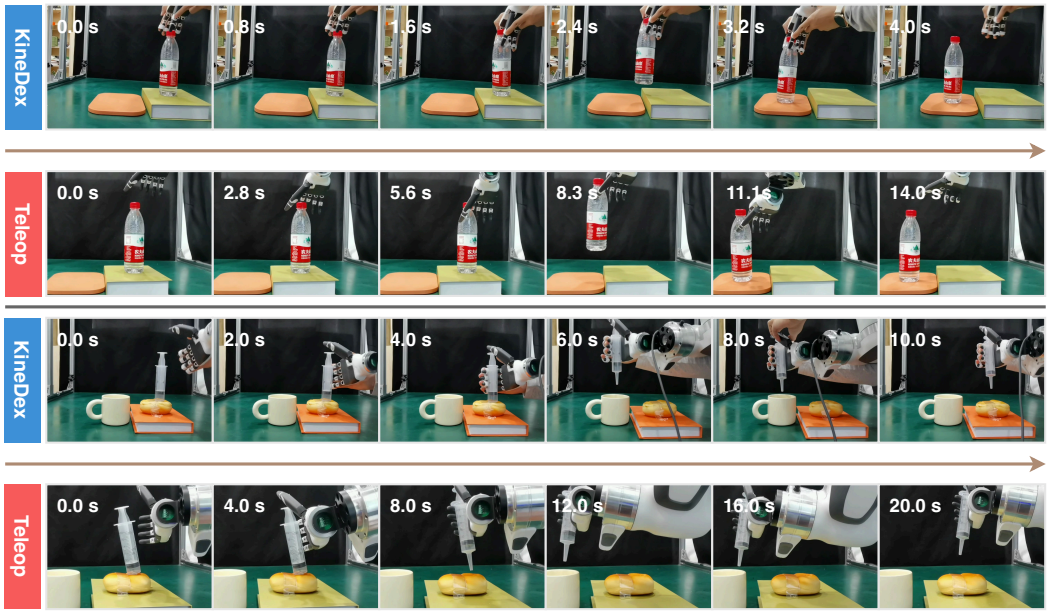


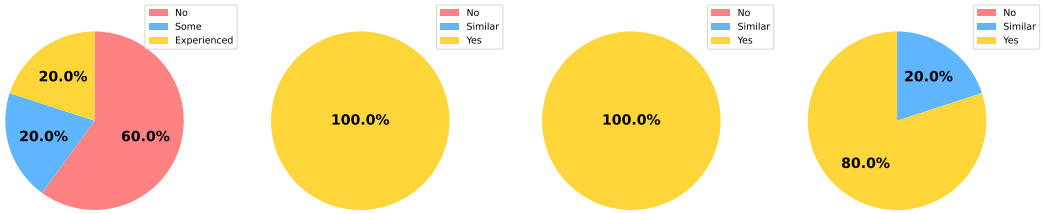
Figure 4: Comparison of demonstration collection time between **KineDex** and teleoperation on the *Bottle Picking* and *Syringe Pressing*.

final three, more challenging tasks, the average success rates still exceed 50%. This drop is likely due to the increased demands for fine-grained localization and contact reasoning, which may exceed the representational capacity of the current policy inputs. Despite these challenges, the results demonstrate that kinesthetic demonstrations effectively support visuomotor policy learning across a wide range of daily manipulation tasks, owing to their natural alignment with human behavior and the availability of accurate tactile and force feedback.

The ablation results without force control underscore the importance of incorporating both force measurements during training and force control during inference. In this setting, the average success rate across all tasks drops to just 16.7%, with even simple tasks such as *Bottle Picking* rarely completed successfully. Without force control, the robotic hand often merely contacts the object surface without applying sufficient pressure, leading to frequent failures in contact-rich tasks. To further examine this effect, Figure 3 visualizes the fingertip-object contact forces during inference. KineDex accurately tracks the desired contact forces predicted by the policy, exhibiting similar magnitudes and temporal patterns. In contrast, without force control, the executed force remains flat and fails to follow the predicted targets, indicating a breakdown in physical interaction quality.

When the policy is trained without tactile input, performance exhibits a moderate drop on most relatively simple pick-and-place tasks. This outcome is expected, as visual information alone is often sufficient to estimate required forces in many scenarios. However, for more contact-intensive tasks such as *Cap Twisting*, *Toothpaste Squeezing*, and *Syringe Pressing*, removing tactile input leads to a significant performance decline, with average success rates decreasing by 26.7%. This suggests that in scenarios with severe visual occlusion or tasks heavily reliant on contact feedback, tactile sensing serves as an effective auxiliary modality that substantially improves task success.

The variant without inpainting yields a zero success rate across all tasks and exhibits unreasonable behaviors during execution. These results confirm that raw kinesthetic demonstrations are infeasible



(a) Do you have experience with teleoperation?

(b) Does **KineDex** help collect more accurate tactile data?

(c) Does **KineDex** help collect demonstrations for more complex tasks?

(d) Is **KineDex** easier to use?

Figure 5: Summary of user study results. Five participants used both the teleoperation system and **KineDex** to collect demonstrations. Pie charts summarize their feedback on key evaluation criteria.

for direct policy training due to out-of-distribution visual observations, and that inpainting provides an effective strategy for mitigating this issue.

4.2 Efficiency Evaluation

We further validate the advantages of KineDex over teleoperation for data collection through comparative experiments. We replicate the Open-TeleVision [15] setup to construct a teleoperation system using a Franka Emika Panda arm equipped with an Inspire Hand[4], and track the operator's hand motion using the Meta Quest 3 headset[5]. To ensure a fair comparison, we implement KineDex using the same Inspire Hand in this set of experiments. Detailed setup is provided in Appendix C.1.

We evaluate five tasks that are feasible under teleoperation and report the demonstration success rates during data collection. As shown in Table 2, teleoperation achieves an average success rate of 39%, whereas KineDex consistently achieves near-perfect success across all tasks. The absence of real-time tactile feedback significantly impairs teleoperation performance, particularly in tasks such as *Cup Picking*, where the operator frequently crushes the paper cup due to excessive force. Moreover, because the operator views the scene through a remote camera in the VR interface, the resulting unnatural visual feedback introduces additional challenges, especially for fine-grained manipulation such as *Syringe Pressing*. These results suggest that teleoperation requires greater operator expertise and repeated trial-and-error to produce high-quality demonstrations, leading to significantly lower data collection efficiency compared to KineDex.

In addition to success rates, we also measure the time required to collect demonstrations with both methods. As shown in Figure 4, the improvement in efficiency is substantial: on the complex task of *Syringe Pressing*, KineDex completes each demonstration in roughly half the time required by teleoperation; on the simpler task of *Bottle Picking*, it takes less than one-third of the time. This efficiency gap is primarily due to the additional time required for the operator to adjust to precise hand poses, as well as the inherent latency and limited responsiveness of the teleoperation system.

4.3 User Study

To provide a more comprehensive evaluation of KineDex, we conduct a user study using the setup described in Appendix C.2. Participants used both KineDex and the teleoperation system, and subsequently provided feedback on their perceived effectiveness and ease of use. The results, summarized in Figure 5, indicate that KineDex outperforms teleoperation across all evaluation criteria. Specifically, all participants agreed that KineDex enables more accurate tactile data collection and is better suited for complex manipulation tasks, while most found it easier to use than teleoperation.

[4] https://www.inspire-robots.com/product/frwz/

[5] https://www.meta.com/quest/quest-3/

5 Conclusion

We present **KineDex**, a framework for collecting tactile-enriched demonstrations and training visuomotor policies for dexterous manipulation. Through experiments on nine contact-rich manipulation tasks, we demonstrate the feasibility of kinesthetic teaching for data collection, and highlight the critical roles of tactile sensing and force control in addressing complex manipulation challenges. Comparative studies with teleoperation further reveal that our approach offers substantial advantages in both data collection efficiency and user experience. We hope that KineDex provides a new perspective for future research on scalable and effective data collection for dexterous robotic hands.

6 Limitations

Based on experimental results and user study feedback, we summarize the limitations of **KineDex** as follows. We view these limitations as promising directions for future work and hope they inspire further advancements in subsequent research:

- Although our results show that inpainting occluded regions is sufficient for training visuomotor policies, its effectiveness may degrade under more severe occlusions. This issue could potentially be mitigated by fine-tuning the inpainting model on robot-specific data.
- The current setup requires two human hands to control a single dexterous hand due to the morphological mismatch between the robotic and human thumbs, limiting scalability for bimanual demonstrations. Future work could explore more biomimetic hardware designs to enable single-handed kinesthetic teaching.

References

[1] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang. Rotating without seeing: Towards in-hand dexterity through touch, Mar. 2023.

[2] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik. General in-hand object rotation with vision and touch, Sept. 2023.

[3] H. Zhang, Z. Wu, L. Huang, S. Christen, and J. Song. Robustdexgrasp: Robust dexterous grasping of general objects from single-view perception, 2025. URL https://arxiv.org/abs/2504.05287.

[4] H. Lee, Y. Kim, V. M. Staven, and C. Sloth. Trajectory optimization for in-hand manipulation with tactile force control, 2025. URL https://arxiv.org/abs/2503.08222.

[5] K. Shaw, A. Agarwal, and D. Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning, Sept. 2023.

[6] K. Shaw and D. Pathak. Leap hand v2: Dexterous, low-cost anthropomorphic hybrid rigid soft hand for robot learning. In *2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS)*, 2024.

[7] C. C. Christoph, M. Eberlein, F. Katsimalis, A. Roberti, A. Sympetheros, M. R. Vogt, D. Liconti, C. Yang, B. G. Cangan, R. J. Hinchet, and R. K. Katzschmann. Orca: An open-source, reliable, cost-effective, anthropomorphic robotic hand for uninterrupted dexterous task learning, Apr. 2025.

[8] B. Romero, H.-S. Fang, P. Agrawal, and E. Adelson. Eyesight hand: Design of a fully-actuated dexterous robot hand with integrated vision-based tactile sensors and compliant actuation, Aug. 2024.

[9] R. Bhirangi, V. Pattabiraman, E. Erciyes, Y. Cao, T. Hellebrekers, and L. Pinto. Anyskin: Plug-and-play skin sensing for robotic touch. *arXiv preprint arXiv:2409.08276*, 2024.

[10] C. Lin, Z. Lin, S. Wang, and H. Xu. Dtact: A vision-based tactile sensor that measures high-resolution 3d geometry directly from darkness, Sept. 2022.

[11] J. Xu, L. Wu, C. Lin, D. Zhao, and H. Xu. Dtactive: A vision-based tactile sensor with active surface, Oct. 2024.

[12] S. Wang, Y. She, B. Romero, and E. Adelson. Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger, June 2021.

[13] M. Gallipoli, S. Buonocore, M. Selvaggio, G. A. Fontanelli, S. Grazioso, and G. Di Gironimo. A virtual reality-based dual-mode robot teleoperation architecture. *Robotica*, 42(6): 1935–1958, June 2024. ISSN 0263-5747, 1469-8668. doi:10.1017/S0263574724000663.

[14] 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, July 2024.

[15] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback, July 2024.

[16] Z. Si, K. L. Zhang, Z. Temel, and O. Kroemer. Tilde: Teleoperation for dexterous in-hand manipulation learning with a deltahand, Aug. 2024.

[17] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators, July 2024.

[18] K. Shaw, S. Bahl, and D. Pathak. Videodex: Learning dexterity from internet videos. In *Proceedings of The 6th Conference on Robot Learning*, pages 654–665. PMLR, Mar. 2023.

[19] H. G. Singh, A. Loquercio, C. Sferrazza, J. Wu, H. Qi, P. Abbeel, and J. Malik. Hand-object interaction pretraining from videos, Sept. 2024.

[20] I. Guzey, Y. Dai, G. Savva, R. Bhirangi, and L. Pinto. Bridging the human to robot dexterity gap through object-oriented rewards, Oct. 2024.

[21] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation, Oct. 2024.

[22] Z. Chen, S. Chen, E. Arlaud, I. Laptev, and C. Schmid. Vividex: Learning vision-based dexterous manipulation from human videos, Sept. 2024.

[23] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation, 2024. URL https://arxiv.org/abs/2410.11792.

[24] H. Xu, M. Chen, G. Li, L. Wei, S. Peng, H. Xu, and Q. Li. An immersive virtual reality bimanual telerobotic system with haptic feedback, Jan. 2025.

[25] H. Zhang, S. Hu, Z. Yuan, and H. Xu. Doglove: Dexterous manipulation with a low-cost open-source haptic force feedback glove, Feb. 2025.

[26] C. Piazza, G. Grioli, M. G. Catalano, and A. Bicchi. A century of robotic hands. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):1–32, 2019.

[27] U. Yoo, J. Francis, J. Oh, and J. Ichnowski. Kinesoft: Learning proprioceptive manipulation policies with soft robot hands, Mar. 2025.

[28] C. Chen, Z. Yu, H. Choi, M. Cutkosky, and J. Bohg. Dexforce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation, Jan. 2025.

[29] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation, Oct. 2024.

[30] S. Zhou, C. Li, K. C. K. Chan, and C. C. Loy. Propainter: Improving propagation and transformer for video inpainting, Sept. 2023.

[31] Y. Qin, W. Yang, B. Huang, K. V. Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system, May 2024.

[32] A. Handa, K. V. 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, Oct. 2019.

[33] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto. Open teach: A versatile teleoperation system for robotic manipulation, Mar. 2024.

[34] S. P. Arunachalam, I. Güzey, S. Chintala, and L. Pinto. Holo-dex: Teaching dexterity with immersive mixed reality, Oct. 2022.

[35] 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, 2024. URL https://arxiv.org/abs/2407.03162.

[36] 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, Mar. 2025.

[37] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, S. Feng, B. Burchfiel, and S. Song. Adaptive compliance policy: Learning approximate compliance for diffusion guided control, Oct. 2024.

[38] T. Ablett, O. Limoyo, A. Sigal, A. Jilani, J. Kelly, K. Siddiqi, F. Hogan, and G. Dudek. Multimodal and force-matched imitation learning with a see-through visuotactile sensor. *IEEE Transactions on Robotics*, 41:946–959, 2025. ISSN 1552-3098, 1941-0468. doi: 10.1109/TRO.2024.3521864.

[39] D. Wei and H. Xu. A wearable robotic hand for hand-over-hand imitation learning, Sept. 2023.

[40] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei. Learning to generalize across long-horizon tasks from human demonstrations, June 2021.

[41] M. Chang and S. Gupta. One-shot visual imitation via attributed waypoints and demonstration augmentation, 2023. URL https://arxiv.org/abs/2302.04856.

[42] T. Yu, P. Abbeel, S. Levine, and C. Finn. One-shot composition of vision-based skills from demonstration. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2643–2650. IEEE, 2019.

[43] E. Johns. Coarse-to-fine imitation learning: Robot manipulation from a single demonstration, 2021. URL https://arxiv.org/abs/2105.06411.

[44] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion, Mar. 2024.

[45] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations, June 2024.

[46] Z. Sun, Z. Shi, J. Chen, Q. Liu, Y. Cui, Q. Ye, and J. Chen. Vtao-bimanip: Masked visual-tactile-action pre-training with object understanding for bimanual dexterous manipulation, Jan. 2025.

[47] I. Guzey, B. Evans, S. Chintala, and L. Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play, Mar. 2023.

[48] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto. See to touch: Learning tactile dexterity through visual incentives, Sept. 2023.

[49] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning, Sept. 2024.

[50] M. Lepert, J. Fang, and J. Bohg. Phantom: Training robots without robots using only human videos, 2025. URL https://arxiv.org/abs/2503.00779.

[51] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, Jan. 2024.

[52] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware, Apr. 2023.

[53] B. Siciliano, O. Khatib, and T. Kröger. *Springer handbook of robotics*, volume 200. Springer, 2008.

[54] M. A. Johnson and M. H. Moradi. *PID control*. Springer, 2005.

[55] J. Carpentier, G. Saurel, G. Buondonno, J. Mirabel, F. Lamiraux, O. Stasse, and N. Mansard. The pinocchio c++ library: A fast and flexible implementation of rigid body dynamics algorithms and their analytical derivatives. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pages 614–619. IEEE, 2019.

[56] J. Carpentier, F. Valenza, N. Mansard, et al. Pinocchio: fast forward and inverse dynamics for poly-articulated systems, 2015–2018. *URL https://stack-of-tasks. github. io/pinocchio*.

A Task Design

We design a suite of nine contact-rich dexterous manipulation tasks to comprehensively evaluate policy performance across diverse interaction modes. In all tasks, object poses are randomized to introduce variability in spatial configurations and contact conditions. The tasks are as follows:

- *Bottle Picking*: Pick up a partially filled plastic water bottle (approximately 200g), transport it to a designated target location, and place it without dropping. This task requires stable control over a compliant object with shifting internal mass.
- *Cup Picking*: Pick up a disposable paper cup, transport it stably, and place it at a specified location. The task emphasizes gentle contact and manipulation of deformable, lightweight objects.
- *Egg Picking*: Pick up a raw egg, move it to a target location, and place it safely. This task demands extremely fine force modulation to prevent cracking or dropping during manipulation.
- *Cap Twisting*: Grasp a plastic bottle, unscrew the cap, and lift it off. This task involves precise torque generation, in-hand stabilization, and coordinated finger-thumb rotation.
- *Nut Tightening*: Rotate a plastic nut clockwise to securely fasten it onto a bolt. The task requires simultaneous rotational motion and downward force application.
- *Peg Insertion*: Insert four wooden pegs into corresponding holes on a board. This evaluates spatial alignment, contact control, and force-guided insertion under tight tolerances.
- *Charger Plugging*: Plug a two-prong charger into a power strip, requiring fine spatial alignment and precise force control to achieve successful insertion.
- *Toothpaste Squeezing*: Flip open the toothpaste cap using the thumb, then squeeze paste onto a toothbrush. The task combines sequential action planning and fine-grained variable force modulation.
- *Syringe Pressing*: Hold a syringe and press the plunger with the thumb to expel water. This simulates a one-handed, force-controlled manipulation requiring steady actuation and grasp stability.

To provide a more intuitive understanding of the tasks, we visualize the execution of the trained policies for all nine contact-rich manipulation tasks in Figure 9. Note that the execution times annotated in the figure may differ from those shown in Figure 4, as the policies are trained with different sets of demonstrations.

B Policy Training Details

Our policy implementation builds upon the official Diffusion Policy codebase[6]. We retain the original architecture of the U-Net-based diffusion model and the multi-view visual encoder without modifications. However, to better support our contact-rich manipulation setting, we introduce the following enhancements:

- We integrate a tactile encoder to process high-resolution tactile inputs from five fingers. The input is a tensor of shape 5x120x3, where each of the five fingers has 120 tactile points, and each point encodes a 3D vector representing the magnitude and direction of contact force. Each finger's data is processed through a shared 1D convolutional encoder to extract per-finger features. These features are then concatenated and passed through a two-layer multilayer perceptron (MLP) to produce a fixed-length tactile embedding. This embedding is fused with visual and proprioceptive observations to form the policy input.
- Our policy outputs a 23-dimensional action vector, comprising 6 degrees of freedom for the end-effector pose, 12 joint angles for the dexterous hand, and 5 normal force targets at the fingertips.

[6] https://github.com/real-stanford/diffusion_policy

To improve control smoothness and temporal consistency, we adopt an action chunking strategy [52] with a chunk length of 16. During execution, we employ an interpolation controller that applies control commands at 100 Hz for the robotic arm and 50 Hz for the dexterous hand, ensuring high-frequency and stable control for both subsystems.

- For relatively simple tasks such as picking, plugging, and insertion, we collect approximately 100 demonstrations per task. For more challenging tasks such as *toothpaste squeezing* and *syringe pressing*, we collect around 150 demonstrations. All baselines are trained for 500 epochs on each task.

Some key hyperparameters for training and inference are summarized in Table 3.

Table 3: Training and inference configuration.

Config	**Value**
Observation horizon	2
Action horizon	16
Observation resolution	240×320
Optimizer	AdamW
Optimizer momentum	$\beta_1, \beta_2 = 0.95, 0.999$
Learning rate	1e-4
Batch size	64
Inference denoising iterations	16
Temporal ensemble steps	8
Temporal ensemble adaptation rate	-0.01

C Experiment Setup Details

C.1 Teleoperation System Setup

We replicate the Open-TeleVision [15] setup and construct a single-handed teleoperation system for comparative evaluation. Our setup consists of the following components:

- **Robot Platform**: A 7-DoF Franka Emika Panda robotic arm mounted with a 6-DoF Inspire Hand.
- **Operator Interface**: A Meta Quest 3 headset is used to track the operator's head motions in real time using its built-in hand tracking system.
- **Motion Retargeting**: The operator's wrist pose is mapped to the robot arm's end-effector via a closed-loop inverse kinematics (CLIK) controller implemented with the Pinocchio library [55, 56], ensuring stable and precise arm control. Simultaneously, the operator's hand keypoints are retargeted to the 12-DoF Inspire Hand using the dex-retargeting framework [31], which optimizes the alignment between human and robot keypoint vectors while maintaining temporal consistency. This unified retargeting approach enables intuitive and responsive control of both the arm and the dexterous hand.

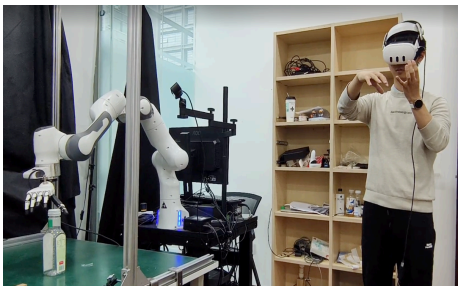
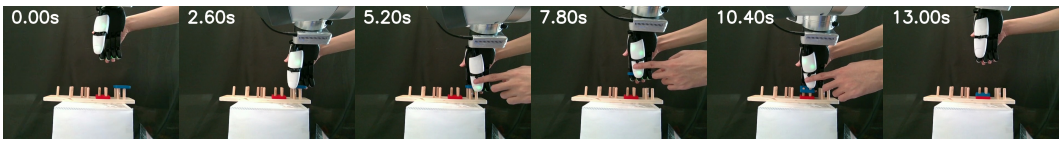
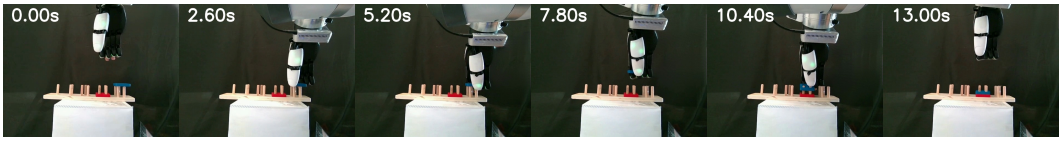


Figure 6: The overview of the teleoperation system setup.



(a) Kinesthetic Demonstration

(b) Masked Demonstration.

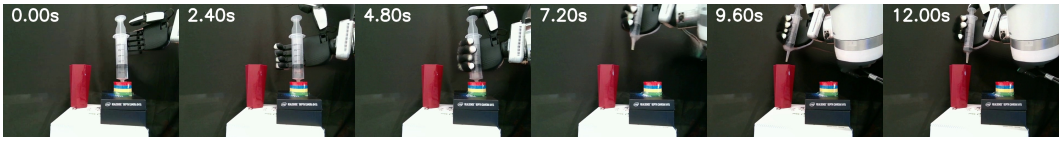


(c) Inpainted Demonstration.

Figure 7: Data preprocessing pipeline for *Peg Insertion*.

(a) Kinesthetic Demonstration

(b) Masked Demonstration.



(c) Inpainted Demonstration.

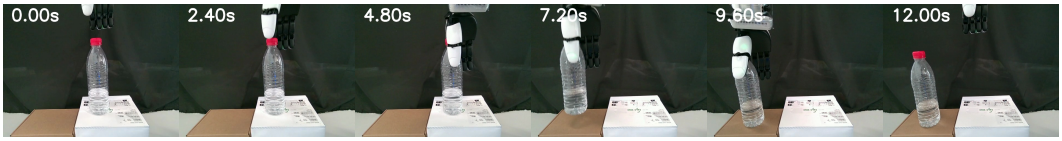
Figure 8: Data preprocessing pipeline for *Syringe Pressing*.

C.2 User Study Setup

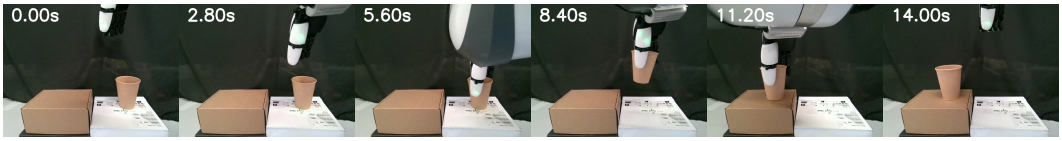
We invited five participants with prior experience in robotics projects to take part in the user study. The participants had varying levels of teleoperation expertise. Each participant was guided to use both KineDex and our custom-built teleoperation system to collect demonstrations. They completed five trials on two tasks of different difficulty levels: *Bottle Picking* and *Syringe Pressing*. Afterward, they evaluated both systems in terms of perceived effectiveness and ease of use.

D Data Preprocessing Details

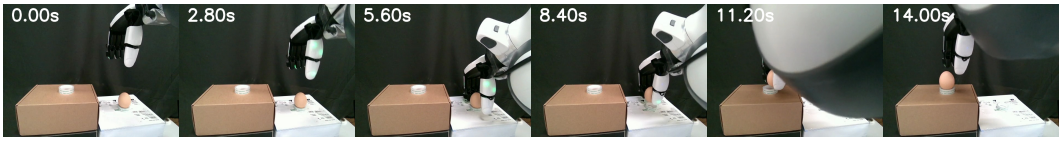
The KineDex data preprocessing pipeline consists of three stages. First, demonstrations are collected via kinesthetic teaching. Second, we use Grounded-SAM [51] to segment human body regions and generate corresponding masks. Finally, we use ProPainter [30] to inpaint the occluded areas by removing the masked regions. Two examples are illustrated in Figures 7 and 8.



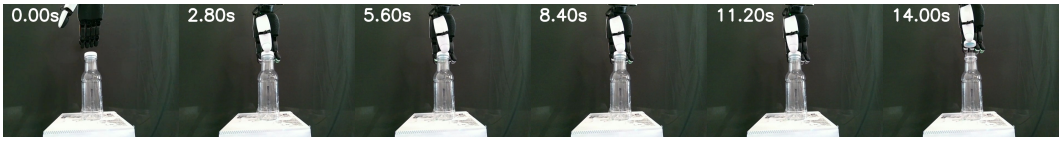
(a) Bottle Picking



(b) Cup Picking

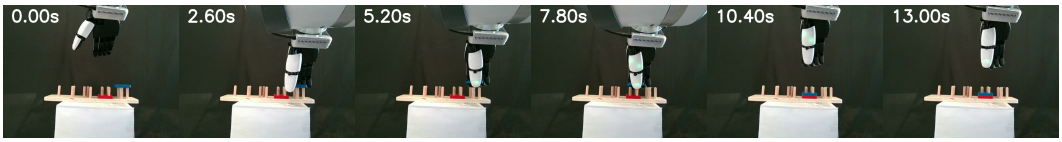


(c) Egg Picking

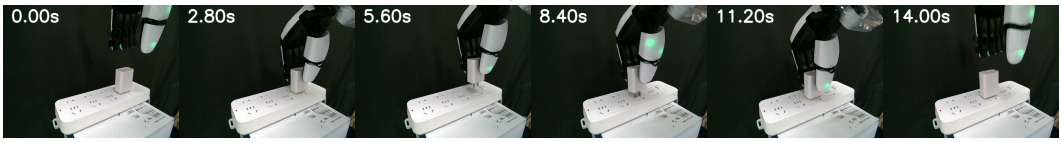


(d) Cap Twisting

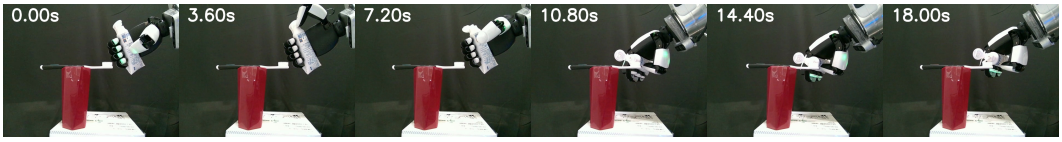
(e) Nut Tightening



(f) Peg Insertion



(g) Charger Plugging



(h) Toothpaste Squeezing

(i) Syringe Pressing

Figure 9: Executions of trained policies on nine contact-rich manipulation tasks.