2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

FreeTacMan: Robot-free Visuo-Tactile Data Collection System for Contact-rich Manipulation

Longyan Wu [1,4*] **Checheng Yu** [2*] **Jieji Ren** [3*] **Li Chen** [2] **Yufei Jiang** [5]
Ran Huang [4] **Guoying Gu** [3] **Hongyang Li** [2,1]
[1] Shanghai Innovation Institute [2] The University of Hong Kong
[3] Shanghai Jiao Tong University [4] Fudan University [5] Shanghai University

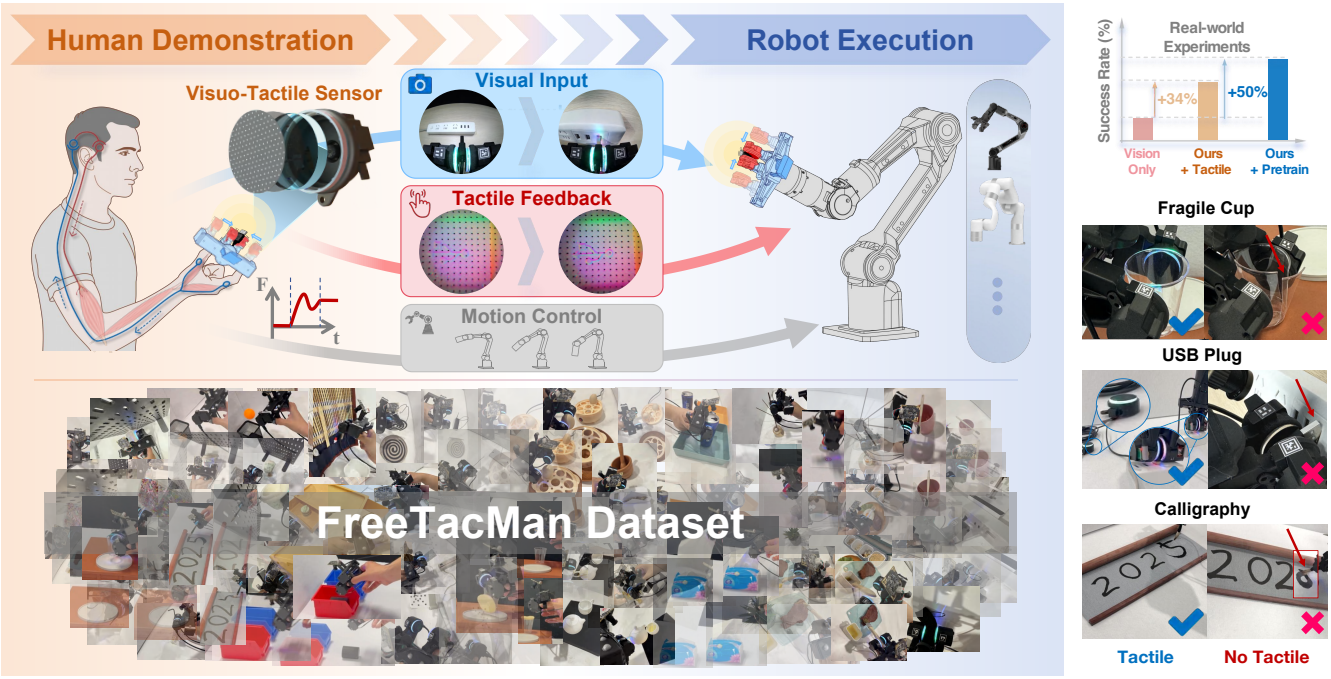

Fig. 1: **Overview of FreeTacMan**. FreeTacMan is a robot-free, human-centric visuo-tactile data collection system that enables the efficient transfer of human visual, tactile, and motor skills to robots. It facilitates the collection of large-scale, contact-rich manipulation datasets. Project page: https://opendrivelab.com/FreeTacMan.

***Abstract*— Enabling robots with contact-rich manipulation remains a pivotal challenge in robot learning, which is substantially hindered by the data collection gap, including its inefficiency and limited sensor setup. While prior work has explored handheld paradigms, their rod-based mechanical structures remain rigid and unintuitive, providing limited tactile feedback and posing challenges for operators. Motivated by the dexterity and force feedback of human motion, we propose FreeTacMan, a human-centric and robot-free data collection system for accurate and efficient robot manipulation. Concretely, we design a wearable gripper with visuo-tactile sensors for data collection, which can be worn by human fingers for intuitive control. A high-precision optical tracking system is introduced to capture end-effector poses while synchronizing visual and tactile feedback simultaneously. We leverage FreeTacMan to collect a large-scale multimodal dataset, comprising over 3000k paired visuo–tactile images with end-effector poses, 10k demonstration trajectories across 50 diverse contact-rich manipulation tasks. FreeTacMan achieves multiple improvements in data collection performance over prior works and enables effective policy learning from self-collected datasets. By open-sourcing the hardware and the dataset, we aim to facilitate reproducibility and support research in visuo-tactile manipulation.**

I. INTRODUCTION

Humans inherently rely on the integration of vision and touch to perform contact-rich manipulation tasks. While vision provides comprehensive object recognition and pose estimation capabilities, tactile feedback conveys critical information about local contacts that cannot be obtained visually, such as surface texture [1], in-hand pose [2], material compliance [3], and force distribution [4]. For example, when handling a fragile or deformable object, vision guides initial motion planning and object localization, while tactile information enables modulation of grip force and adjustment of object orientation to prevent damage.

The vision-based imitation learning has shown strong potential in robot manipulation tasks [5], [6], [7], [8], benefiting from the growing large-scale demonstration datasets [9], [10]. In the tactile domain, visuo-tactile sensors offer high-resolution, sensitive, and easily integrable multimodal signals, making them particularly well-suited for existing policy learning pipelines. However, the lack of large-scale, high-

quality tactile datasets and compatible sensing hardware has prevented similar advances. To accelerate research in this domain, two prerequisites would be addressed: 1) a visuo-tactile data collection system that provides real-time tactile feedback and enables rapid redeployment, and 2) a large-scale, high-precision visuo-tactile dataset covering diverse contact-rich manipulation tasks [11].

Early efforts concentrated on collecting data using sensors mounted on robots [12], [13], which limit the flexibility for scaling visuo-tactile data. Systems that rely on motion-capture with AR/VR visualization [14] or primary-replica teleoperation rigs [15] capture accurate camera views and trajectories. They conversely offer no direct, real-time tactile signals, and impose a fixed robot setup with complex calibration or high latency. Handheld data collection paradigms [16], [17] release the human operators from robot embodiment. However, these methods typically rely on SLAM and IMU fusion for localization, which introduces significant positioning errors that are particularly detrimental to tactile perception. Moreover, they transmit tactile cues through long mechanical linkages, hindering direct tactile feedback and precise measurement of instantaneous tactile signals.

The inefficiency and lack of real-time tactile feedback in current data collection setups compromise data quality, limit scalability, and pose a risk of sensor damage during operation. Consequently, existing visuo-tactile datasets[18], [19] are largely confined to specific perception tasks or lack the diversity and precision required for generalizable visuo-tactile policy learning across diverse manipulation scenarios.

In this work, we introduce **FreeTacMan**, a robot-free and human-centric visuo-tactile data collection system to acquire manipulation data accurately and efficiently. With a modular sensor and in-situ[1] hardware, it ensures precise, real-time tactile feedback via two mechanisms - a gripper-finger interface coupled to human fingertips, and a linear transmission mechanism enabling precise gripper control and unattenuated tactile feedback. This foundation enables a comprehensive visuo-tactile manipulation dataset with high precision, diverse task scenarios, and rich contact dynamics, recorded with synchronized high-resolution visual, visuo-tactile, and pose data. This extensive collection provides a robust foundation for training and evaluating generalizable visuo-tactile policies. To validate the effectiveness of the data collected by FreeTacMan, we train and deploy imitation learning policies that leverage a temporal-aware tactile pretraining strategy on challenging manipulation tasks.

In summary, our main **contributions** are:

(i) An in-situ, robot-free, real-time tactile data-collection *system* that leverages a handheld gripper with modular visuo-tactile sensors to excel at diverse contact-rich tasks efficiently.

(ii) A large-scale, high-precision (sub-millimeter) visuo-tactile manipulation *dataset* with over 3000k visuo-tactile image pairs, more than 10k trajectories across 50 tasks.

(iii) Experimental validation shows that imitation policies trained with our visuo-tactile data achieve an average 50% higher success rate than vision-only approaches across challenging contact-rich manipulation tasks.

[1]"In-situ" indicates that FreeTacMan preserves natural fingertip–environment interaction. It emphasizes maintaining the original grasp posture while capturing tactile feedback. This is akin to the concept in materials science [20] and biology [21], where systems are studied in their native conditions—without altering their environment or state—to gain a deeper understanding of their behavior under actual working conditions.

TABLE I: **Comparison with existing data collection systems.** Our in-situ design delivers direct and high-precision tactile feedback to operators. To evaluate tactile feedback fidelity of handheld methods quantitatively, we measure the number of mechanical transmission "link" inside the handheld gripper from the human hand to the grasped object.

Category	Method	Control Method	Tactile Feedback
Teleop: VR/AR	ARCap [22]	VR Controller	–
	DexCap [23]	Hand Mocap	–
	TactAR [14]	VR Controller	Visual
	Bunny-VisionPro [24]	Hand Retargeting	Vibration
Teleop: Primary–Replica	ALOHA [15]	Puppet Arm	–
	GELLO [25]	Puppet Arm	–
	Bi-ACT [26]	Puppet Arm	Force
Handheld	UMI [16], FastUMI [17] ViTaMIn [27]	Trigger	Contact (4 links)
In-situ	**FreeTacMan (Ours)**	Fingertips	Touch (1 link)

II. Related Work

A. Dataset Collection System for Robot Learning.

Recent robot learning has been greatly advanced through imitating expert demonstrations [28]. Popular approaches [29], [30] leverage large-scale visuomotor datasets (pairs of RGBD sensors and actions) to train end-to-end policies that generalize across objects and scenes. To gather data suitable for imitation learning, prior works have relied on different interfaces to teleoperate real robots, including motion capture with AR or VR visualization [22], [10], [14], 3D space mouse [30], primary-replica system [15], [25], [26], and wearable devices [31]. However, these setups are either expensive, operationally complex, or suffer from limited precision. Handheld data collection paradigms [16] offer greater flexibility for various robot embodiments, but their trigger-based grippers and multi-link transmission designs introduce backlash between links, which blurs tactile cues and leads to inaccurate relay of gripper contact to the operator. In contrast, FreeTacMan implements an in-situ data collection system that enables operators to feel gripper contact directly in real time. The comparison of control methods and tactile feedback between ours and prior work is depicted in Table I.

B. Tactile dataset for contact-rich Manipulation.

Most existing visuo-tactile datasets focus on perception tasks such as 6DoF pose tracking [32], cross-modal generation [18], representation learning [19], garment-feature spatially-aligned perception [33], and tactile-semantic description [34], with limited coverage of the full manipulation process. The ObjectFolder benchmark [35] includes 4 visuo-tactile manipulation tasks, exhibiting limited scale and diversity. Moreover, most tactile manipulation datasets are collected with tactile array sensors based on piezoresistive

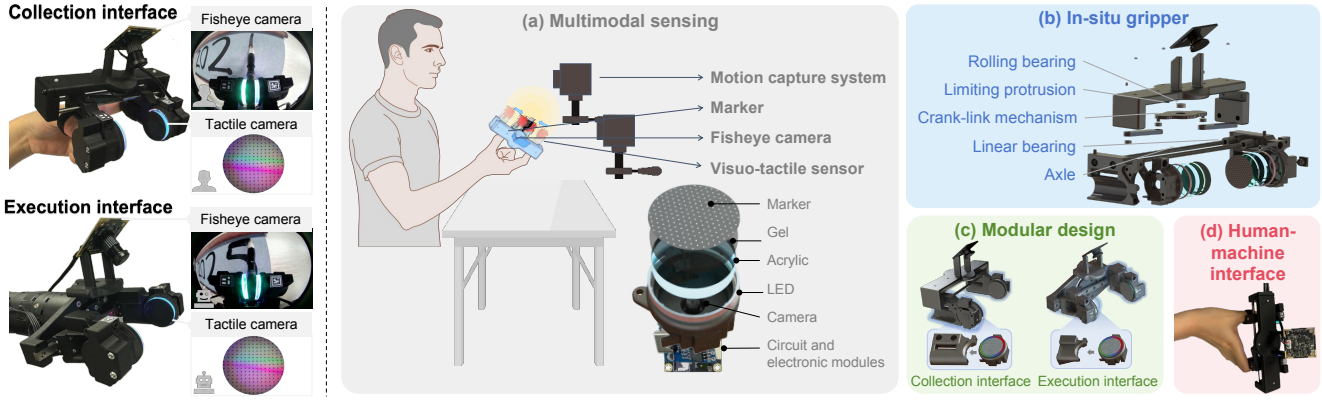

Fig. 2: **Hardware system**. *Left*: The in-situ gripper in the collection and execution interface respectively, with identical visual and tactile observations. *Right*: (a) Composition of the sensor. (b) Exploded view of FreeTacMan. (c) The modular design allows for an agile switch between the collection and execution interface. (d) Human-machine interface design.

or Hall-effect principles [13], [36], [37], [38]. Although policies trained on these datasets show promising results, the intrinsic limitations of these sensors, such as low spatial resolution, crosstalk, complex fabrication, and environment interference, can constrain policy performance and reduce deployment efficiency in real-world settings. Taken together, these limitations expose an urgent demand for a visuo-tactile dataset that offers rich, high-resolution, and scalable data spanning the entire manipulation process.

III. Method

A. Hardware Design

Design criteria. To enable efficient collection of high-fidelity tactile data, we define the following criteria. **(a)** Multimodal data acquisition: The system must provide competitive visuo-tactile sensing with exceptional consistency, coupled with high-precision pose tracking. **(b)** Efficiency: The system should minimize the tactile transmission path from human fingers to the grasped object for real-time and precise tactile feedback, while ensuring stable and fingertip-level control. **(c)** Scalability: A modular design architecture should be adopted to ensure compatibility across robot embodiments. **(d)** Usability: The system should accommodate a wide range of fingertip sizes, providing ergonomic comfort during operation.

Multimodal sensing. As shown in Fig. 2(a), we employ visuo-tactile sensors and a wrist-mounted camera to capture tactile and visual information. End-effector poses are tracked using a high-precision motion capture system, achieving sub-millimeter accuracy—an essential factor for contact-rich tasks, where even minute pose errors can result in disproportionately large deviations in tactile feedback.

In-situ gripper. FreeTacMan achieves hinge-free operation through visuo-tactile sensors mounted on the operator's fingertip, where the sensor layer forms the interface between skin and manipulated objects, eliminating intermediate linkages to provide zero mechanical attenuation and natural proprioception. To ensure that the unattenuated tactile feedback directly translates to operator control precision, as illustrated in Fig. 2(b), the system incorporates a linear transmission mechanism with chrome-plated steel shafts and linear bearings, constraining movement to highly accurate linear trajectories (axial deviation $\geq$ 0.02 mm). Additionally, an inverted crank-slider mechanism converts finger-driven motion into synchronized linear output, while dual parallel shafts and rolling bearings in linking joints minimize friction and lateral torque, achieving over 90% transmission efficiency.

Modular architecture. FreeTacMan system is built as three plug-and-play modules, each optimized for rapid setup and cross-embodiment compatibility: a sensor perception module for tactile data collection, a universal gripper interface (Fig. 2(c)) for robot compatibility, and a camera mounting scaffold to ensure aligned visual feedback from wrist camera. The sensor is based on the McTac design [39] and features enhanced modularity and consistency. The improved universal mechanical interface allows the same sensor unit to be used interchangeably on a data collection setup or a robotic arm. Automated fabrication ensures consistent quality. Customized interfaces are provided for different robot arms, including a 6-DOF Piper arm for low-load tasks and a 7-DOF Franka arm for high-precision, heavy-load applications. For 3D models of end-effectors and their integration on different robot arms (*e.g.*, Piper, Franka), please visit our demo page.

Human-machine interface design. To achieve rapid adaptability, we incorporate hook-and-loop straps for fingertip fastening, as illustrated in Fig. 2(d). These straps fit hand sizes ranging from the 5% to 95% percentile of adults and support repeated use [40], balancing operational efficiency. Furthermore, the system features a lightweight $(157.5, g)$ and compact $(145 \times 85 \times 106, \text{mm}^3)$ design that ensures comfort.

B. Human-to-Robot Data Transfer

To facilitate human-to-robot skill transfer, a high-precision NOKOV Motion Capture System is used for 6D pose tracking of the interface at 240 Hz, achieving sub-millimeter accuracy. Five retro-reflective markers are mounted on the interfaces, with three positioned on the top plate to measure pose and two on the grippers to capture relative displacement. The coordinates of markers are first transformed from the world coordinate frame to the robot base coordinate frame. A local coordinate frame, aligned with the robot URDF, is established

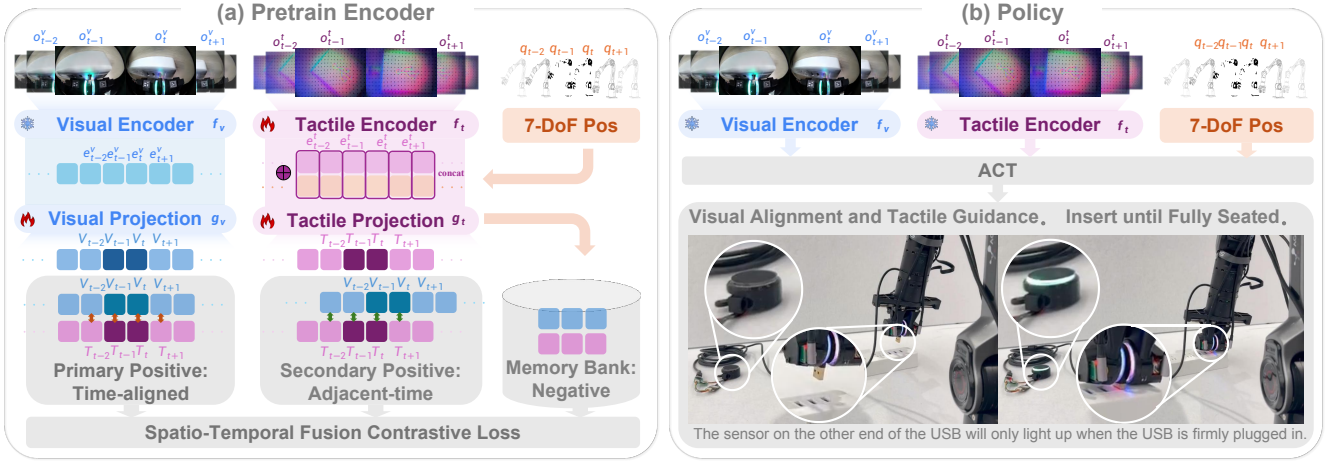

Fig. 3: **Tactile pretraining and policy learning pipeline.** (a) A tactile encoder is pretrained using the self-collected dataset. (b) The pretrained tactile encoder is integrated into an ACT-based policy for downstream tasks such as USB insertion.

at the gripper's Tool Center Point (TCP) using three top-plate markers, allowing derivation of end-effector pose while ensuring consistency with the robot kinematic model. We downsample the tracking data to synchronize with RGB images. To this end, each frame contains: the wrist camera RGB image, two visuo-tactile images, the end-effector pose of the in-situ gripper in the world coordinate frame, and gripper width. Each trajectory consists of a sequence of such embodiment-agnostic data synchronized at 30 Hz.

Unlike prior works [16], [17] relying on SLAM and IMU fusion to estimate end-effector poses, the motion capture system avoids IMU drift and tracking errors. Given the URDF of a target embodiment, we could utilize IKPY [41] as an inverse kinematic solver to map poses of the in-situ gripper to joint positions directly as the action representation.

C. Tactile Pretraining and Policy Learning

A two-stage approach is employed to learn visuo-tactile manipulation policies: 1) Tactile representation learning, 2) visuo-tactile policy learning.

Tactile representation learning. Our wearable system yields high-precision, contact-rich visuo-tactile trajectories, enabling a high-fidelity dataset for representation learning. Although tactile outputs resemble 2D images, applying a vision encoder pretrained on RGB data often produces suboptimal features due to the domain gap in appearance and semantics [42], [27]. We adopt a CLIP-style [43] contrastive pretraining procedure to bridge the domain gap.

As illustrated in Fig. 3(a), both the visual encoder f_v and tactile encoder f_t share a ResNet backbone initialized from the same checkpoint. f_v remains frozen during pretraining, while f_t is finetuned. Each encoder is followed by a projection head: g_v for vision, and g_t for tactile, where g_t first concatenates the tactile features with the normalized 7-DOF joint position vector $\mathbf{q}_i$ to inject robot joint state as global context. At each timestep i, we compute the normalized embeddings $\mathbf{v}_i$ and $\mathbf{t}_i$. We designate $\mathbf{v}_i$ as the *primary positive* for $\mathbf{t}_i$ to align tactile features with the current visual features. However, relying solely on time-aligned contrastive loss neglects temporal dynamics, leading to embeddings that vary abruptly and fail to capture evolving contact patterns. To enforce temporal awareness, a *secondary positive* $\mathbf{v}_{i+1}$ drawn from the next timestep is introduced. All other entries from negatives in a fixed-size memory bank $\mathcal{M}$. We train f_t, g_v and g_t by minimizing the following contrastive loss:

$$L = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{\mathbf{v}_i^\top \mathbf{t}_i/\tau} + e^{\mathbf{v}_{i+1}^\top \mathbf{t}_i/\tau}}{e^{\mathbf{v}_i^\top \mathbf{t}_i/\tau} + e^{\mathbf{v}_{i+1}^\top \mathbf{t}_i/\tau} + \sum_{j\in\mathcal{N}_i} e^{\mathbf{v}_j^\top \mathbf{t}_i/\tau}}, \tag{1}$$

where B indicates batch size, τ is a learned temperature parameter and $\mathcal{N}_i$ indexes the negatives.

Visuo-tactile action chunking transformer. We employ the pretrained tactile encoder to extract tactile representations. As shown in Fig. 3(b), vision and tactile embeddings are then concatenated and input into the action chunking transformer (ACT) [29], which is trained to predict joint positions.

IV. Experiments

We design experiments to answer three key questions:

Q1. Can demonstrations be collected efficiently and accurately using FreeTacMan compared to previous setups?

Q2. Is the in-situ tactile information in FreeTacMan dataset effective for contact-rich tasks policy learning?

Q3. How does the tactile encoder pretrained on self-collected visuo-tactile data improve policy learning?

A. Dataset

Enabled by the efficient, precise, and faithful tactile data collection system, we curate a diverse dataset spanning vision, touch, and proprioception modalities, as illustrated in Fig. 4. The dataset spans 50 tasks, comprising more than 10k trajectories and over 3000k visuo-tactile image pairs. Collected with high-consistent and high-resolution visuo-tactile sensors, it enables large-scale tactile pretraining and multimodal policy learning. Furthermore, the dataset covers diverse fundamental tactile capabilities, as shown in Fig. 4(c). To ensure dataset quality, every trajectory is replayed in our

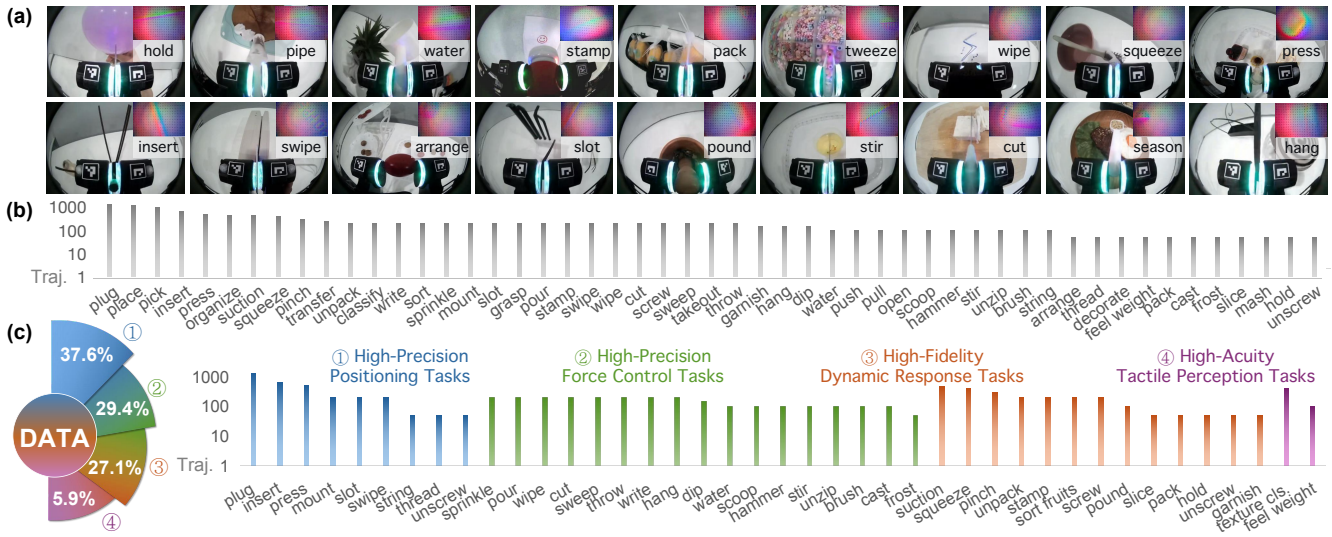

Fig. 4: **The FreeTacMan dataset.** (a) Representative examples illustrating the diversity in both task complexity and tactile context. (b) The dataset covers 50 tasks and features a large-scale collection of data, including more than 10k trajectories and over 3000k visuo-tactile pairs. (c) The dataset enables diverse fundamental tactile capabilities.

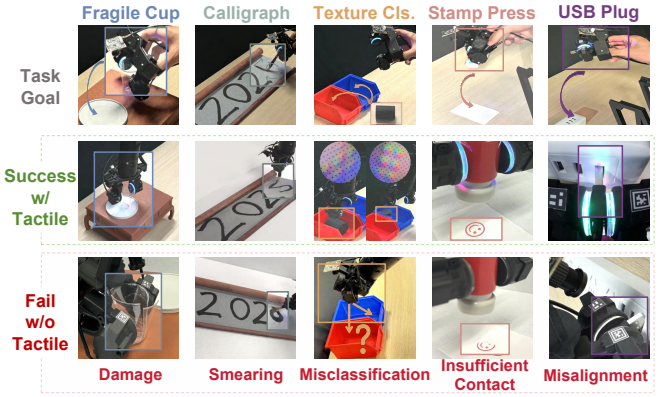

Fig. 5: **Human demonstrations and policy rollouts.** The top row shows goal trajectories, the middle row demonstrates successful rollouts with tactile feedback, and the bottom row showcases typical failure modes without tactile input.

validation process, providing a robust filter that guarantees the reliability of the data for downstream robot learning. The dataset, along with its structure, data format, and license, is available at our dataset page.

B. Experimental Setup

To answer the questions above, we evaluate the effectiveness of FreeTacMan system and the quality of the collected dataset through a diverse set of contact-rich manipulation tasks, shown in Fig. 5. 1) **Fragile cup**. The robot grasps a fragile cup without causing damage and places it stably on a tray. 2) **Calligraphy**. The robot traces the digit "5" with a calligraphy brush. 3) **Texture classification**. The robot grasps and identifies one of two cylindrical objects with distinct textures and sorts it into the correct bin. 4) **Stamp press**. The robot presses a stamp onto paper to produce a clear imprint. 5) **USB plug**. The robot needs to securely insert a pre-grasped USB plug into a socket (error $< 1mm$). These tasks collectively represent the four core tactile capabilities of our dataset: force control (fragile cup), hybrid force-position control (USB plug, stamp press), high-acuity tactile perception (texture classification), and dynamic response (calligraphy).

C. User Study on Data Collection System

Procedure. We evaluate the usability of FreeTacMan through a user study including 12 volunteers, with varying experience in data collection. Besides FreeTacMan, users collect demonstrations using two typical methods: primary–replica-based teleoperation (*i.e.*, ALOHA [15]) and handheld devices (*i.e.*, UMI [16]). To minimize biases, no device-specific instructions are provided and each participant conducts three trials with each device for each task. The participants are instructed to solve tasks as best they can while avoiding collisions and damage. If a task fails, they will continue from where it is interrupted, and the failure will be recorded.

Metrics. We record the task success/failure mode, completion time, and any instances of slippage or damage. Three metrics are adopted to quantify the capability of a particular data collection approach, namely `completion_rate` (fully completed tasks as a percentage of those initiated), collection `efficiency` (the inverse of data collection time), and an overall score, Completion per Unit Time (`CPUT`), defined as `completion_rate` $\times$ `efficiency`.

Demonstrations could be collected efficiently and accurately using FreeTacMan compared to previous setups (Q1). The results of user study are shown in Fig. 6, where FreeTacMan consistently yields the top completion rate and efficiency. CPUT score in Fig. 6(c) highlights the overall advantage of FreeTacMan, gaining 5.05$\times$ higher performance than teleoperation and 1.52$\times$ higher than UMI.

For simple tasks like texture classification that don't require accurate force sensing or control, our system and UMI achieve comparable completion rates, while ALOHA performs slightly inferior. In completion time, we have a 2.65$\times$ advantage

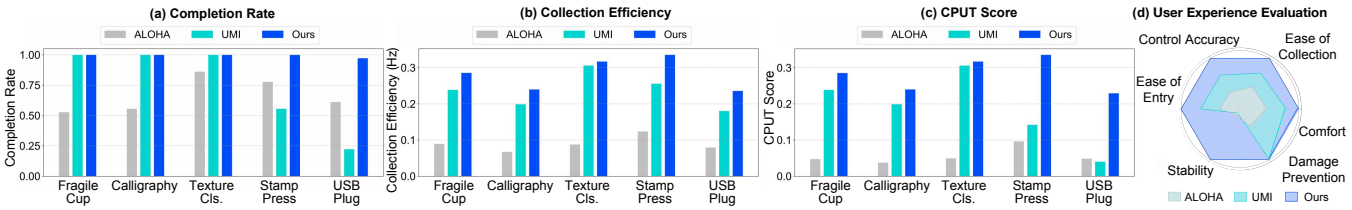

Fig. 6: **User Study on data collection.** (a-c) FreeTacMan outperforms ALOHA [15] and UMI [16] in terms of completion rate, collection efficiency, and the CPUT score per task. (d) FreeTacMan excels at user experience evaluation as well.

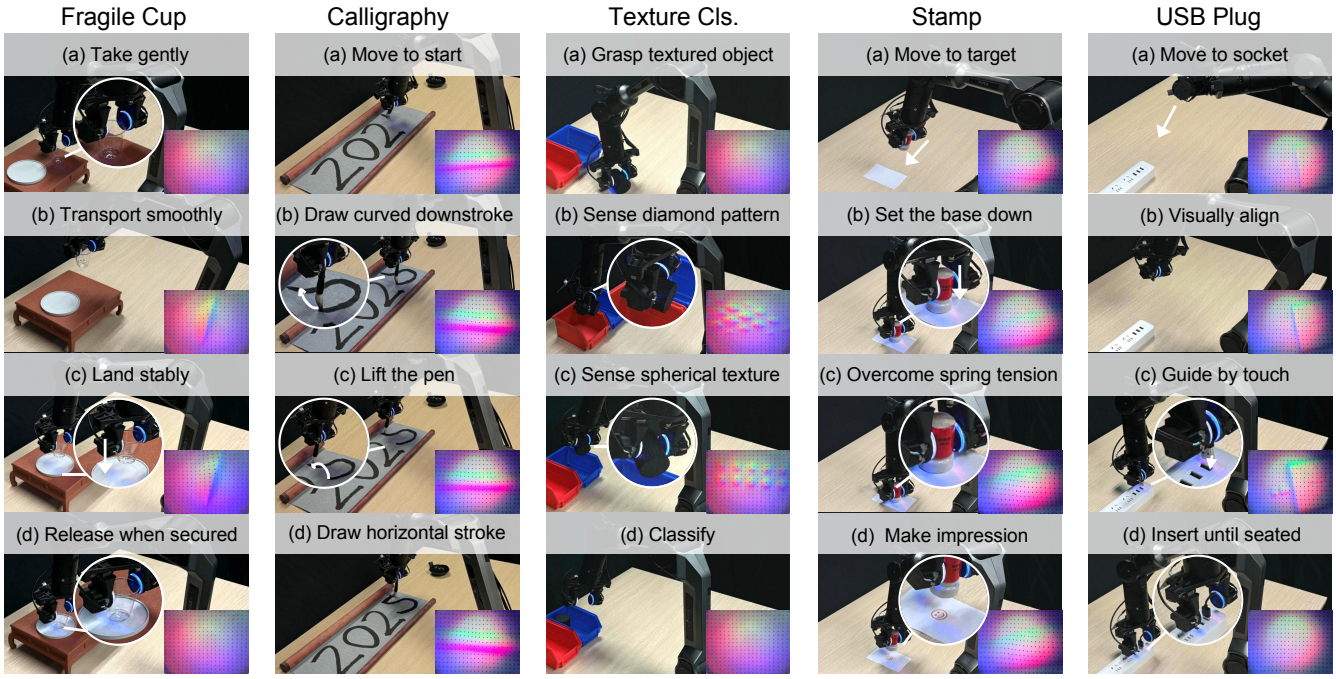

Fig. 7: **Trajectory visualization.** We test FreeTacMan on a variety of contact-rich tasks. Videos are available on our website.

over ALOHA and a 1.22× advantage over UMI. In more complex tasks, such as fragile cup manipulation, where force feedback is necessary to ensure successful grasping and prevent excessive force, primary-replica teleoperation leads to damage-related failures. This suggests that the handheld method provides better force feedback for delicate objects.

For tasks requiring dynamic force control, such as USB plug and stamp press, ALOHA relies on brute force, barely completing them at the risk of causing damage, while UMI often fails due to the lack of slippage detection. In contrast, FreeTacMan combines both precise force perception and control, resulting in superior accuracy and safety. For high-precision trajectory control (*e.g.*, calligraphy), ALOHA struggles significantly—users often need to manually assist the teaching arm with their other hand, resulting in the longest completion time. While UMI can complete the task, the trajectory smoothness remains inferior to that of FreeTacMan.

Fig. 6(d) summarizes the user experience evaluation. Stability, comfort, ease of collection and entry are assessed via questionnaires and normalized. Stability and damage represent failures of object drops and damage. The results show that FreeTacMan is the most user-friendly and reliable data collection system among the three approaches.

TABLE II: **Policy success rates (%) across tasks.** Tactile input and pretraining significantly boost imitation learning.

Method	Fragile Cup	Calligraphy	Texture Cls.	Stamp Press	USB Plug	**Avg.**
ACT [29] (Vision-only)	35	20	20	30	0	21
Ours (+ Tactile w/o Pretrain)	75	70	55	65	10	55
Ours (+ Pretrain)	**80**	**90**	**85**	**80**	**20**	**71**

D. Validation on Imitation Learning

Each task in Fig. 7 is trained and evaluated over 20 trials.

- **ACT** [29] **(Vision-only)**: The original ACT uses RGB images from the wrist camera as input only.
- **Ours (+ Tactile w/o Pretrain)**: An extended ACT model taking both visual and tactile observations, which are separately encoded by identical backbones without pretraining.
- **Ours (+ Pretrain)**: Our full model, where the tactile encoder is pretrained with a multi-positive contrastive objective incorporating both primary and secondary positives.

Integration of tactile feedback results in significant improvements on policy performance (Q2). As illustrated in Table II, the vision-only baseline achieves low performance

TABLE III: **Comparison in unseen object.** Tactile input enables generalization to unseen objects.

	Vision -only	Ours (+ Tactile w/o Pretrain)	Ours (+ Pretrain)
training object	20%	55%	85%
unseen object	**15%**	**55%**	**70%**

across all tasks, with an average success rate of 21%. Without tactile feedback, the robot struggles in tasks requiring fine tactile perception, such as USB plug, calligraphy, and texture classification. When tactile feedback is incorporated naively, *i.e.*, without pretraining, performance improves substantially, with the success rate increasing to 55%. Significant gains are observed in tasks such as fragile cup manipulation (75%) and calligraphy (70%). This highlights the utility of tactile sensing in contact-rich tasks where visual cues alone are insufficient to distinguish subtle features. However, the performance in tasks like USB plug (10%) and texture classification (55%) still lags behind, indicating room for further refinement.

Tactile inputs enable excellent performance even on unseen objects (Q2). In the texture classification task, we select a texture object with an out-of-distribution color (red) for testing. As shown in Table III, tactile input achieves 55% accuracy on unseen objects, matching performance on training objects and significantly exceeding the 15% vision-only baseline. Pretraining further increases accuracy to 70%, demonstrating robust generalization.

Temporal-aware pretraining improves the performance by aligning visual and tactile embeddings (Q3). Incorporating both time-aligned and time-adjacent pairs in CLIP pretraining boosts the average success rate to 71%. As shown in the last line in Table II, the improvement comes from tasks that demand fine-grained control and the ability to track contact dynamics, such as calligraphy (90%) and stamp (80%). Meaningful gain is witnessed in USB plug (20%). Nevertheless, this 20% success rate highlights the challenge of high-precision tasks, where sub-millimeter deviations lead to failure. This is attributed to limited inverse kinematics accuracy and insufficient modeling of insertion dynamics. Future work will thus explore higher-precision arms and reinforcement learning to better learn contact-rich policies. Overall, these results confirm that our approach aligns tactile and visual embeddings with spatial-temporal context.

Pretraining demonstrates substantial benefits under data-scarce conditions (Q3). As shown in Fig. 8, ACT policy with pretrained encoder achieves superior performance with only 50 task-specific demonstrations—55% for cup manipulation, 50% for calligraphy and 60% for texture classification—compared to the visual-only baseline (35%, 20%, and 20%, respectively). With 100 episodes, its performance in cup and stamp tasks (60% each) nearly matches the non-pretrained tactile model (70% and 65%). We further analyze performance discrepancies across task types. Perception-centric tasks like texture classification benefit from the generalized features of the pre-training, while action-oriented tasks such as cup manipulation and calligraphy require not only robust features but also sufficient task-specific demonstrations to learn effective action strategies.

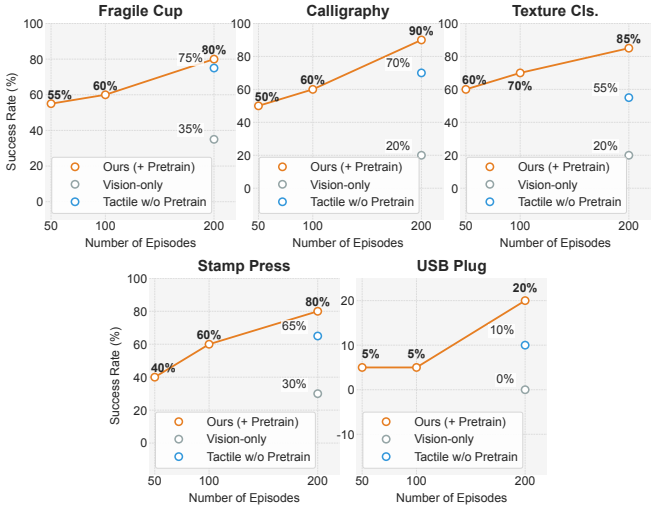

Fig. 8: **Ablation across training episodes.** Our pretrained policy, even with small data, outperforms the vision-only policy trained with large amounts of data. Moreover, it approaches the performance of a large-data tactile-only policy.

TABLE IV: **Cross-sensor generalization results.** Visuo-tactile pretraining helps generalization to other visuo-tactile sensor setups. ID: in-domain. OOD: out-of-domain.

Task	Phase	Ours (+ Tactile w/o Pretrain)		Ours (+ Pretrain)	
		ID Sensor	OOD Sensor	ID Sensor	OOD Sensor
Texture Cls.	Grasp	75%	**80%**	85%	**85%**
	Classify	55%	**40%**	85%	**75%**
	Whole Task	55%	**40%**	85%	**75%**
Fragile Cup	Pick	80%	**70%**	80%	**80%**
	Land	75%	**10%**	80%	**80%**
	Whole Task	75%	**10%**	80%	**80%**

Pretraining on FreeTacMan dataset provides robustness against variations among visuo-tactile sensors (Q3). While our main experiments are conducted in an in-domain setting, where training and evaluation use different sensor batches with high consistency, we also evaluate a more challenging out-of-domain scenario by introducing substantial variations across visuo-tactile sensors. As shown in Table IV, the pretrained model generalizes well across two distinct sensors, which differ in marker angle (0° vs. 45°) and RGB lighting environment (side vs. bottom). The in-domain and out-of-domain performances remain close on the texture classification task (0.85 vs. 0.75) and identical on the fragile cup task (both 0.8). In contrast, the non-pretrained counterpart performs poorly, particularly in phases requiring fine-grained tactile perception: a drop from 80% to 40% in the tactile-based classification phase, and a catastrophic drop from 70% to 10% in the landing phase (which relies on detecting contact with the table before releasing). These results validate the effectiveness of tactile pretraining for cross-sensor generalization.

V. Conclusion

We present FreeTacMan, a human-centric and robot-free data collection system with in-situ visuo-tactile feedback and recording. An in-situ and modular gripper with visuo-tactile sensors enables rapid adaptation, and a large-scale high-precision dataset is collected to support contact-rich manipulation policy learning. Experimental results demonstrate that the proposed system outperforms existing methods in multiple aspects, including data collection efficiency, control accuracy, and human-machine interaction experience. Through policy validation, the effectiveness of the collected data and the importance of visuo-tactile pretraining are further confirmed.

Limitation and future work. While FreeTacMan has demonstrated efficacy across a range of challenging tasks, a few limitations remain. To eliminate dependency on external base stations for submillimeter-level localization, we will develop high-precision visual algorithms for collecting data in the wild. We also plan to extend our system and dataset to dexterous hands and bimanual long-horizon tasks, facilitating study on finer-grained and more complex scenarios.

VI. Acknowledgments

This work is in part supported by the JC STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust. This work was also partially supported by the NSFC Grant No.52505029, and the STCSM Grant No.25ZR1401191 and No. 24511103400.

References

[1] C. E. Connor and K. O. Johnson, "Neural coding of tactile texture: comparison of spatial and temporal mechanisms for roughness perception," *Journal of Neuroscience*, 1992.

[2] W. Xu, Z. Yu, H. Xue, R. Ye, S. Yao, and C. Lu, "Visual-tactile sensing for in-hand object reconstruction," in *CVPR*, 2023.

[3] W. M. B. Tiest and A. M. Kappers, "Cues for haptic perception of compliance," *IEEE Trans. on Haptics*, 2009.

[4] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force estimation using gelslim and inverse fem," in *ICRA*, 2019.

[5] J. Yang, K. Lin, J. Li, W. Zhang, T. Lin, L. Wu, Z. Su, H. Zhao, Y.-Q. Zhang, L. Chen *et al.*, "RISE: Self-improving robot policy with compositional world model," *arXiv preprint arXiv:2602.11075*, 2026.

[6] M. Shi, S. Peng, J. Chen, H. Jiang, Y. Li, D. Huang, P. Luo *et al.*, "EgoHumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration," *arXiv preprint arXiv:2602.10106*, 2026.

[7] M. Shi, L. Chen, J. Chen, Y. Lu, C. Liu, G. Ren, P. Luo, D. Huang, M. Yao, and H. Li, "Is diversity all you need for scalable robotic manipulation?" *TRO*, 2026.

[8] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng *et al.*, "WholeBodyVLA: Towards unified latent vla for whole-body loco-manipulation control," in *ICLR*, 2026.

[9] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *ICRA*, 2024.

[10] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Huang *et al.*, "AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *IROS*, 2025.

[11] L. Chen, C. Sima, K. Chitta, A. Loquercio, P. Luo, Y. Ma, and H. Li, "Intelligent robot manipulation requires self-directed learning," *Authorea Preprints*, 2025.

[12] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," in *ICRA*, 2025.

[13] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3D-ViTac: Learning fine-grained manipulation with visuo-tactile sensing," in *CoRL*, 2024.

[14] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu, "Reactive Diffusion Policy: Slow-fast visual-tactile policy learning for contact-rich manipulation," in *RSS*, 2025.

[15] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *CoRL*, 2024.

[16] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal Manipulation Interface: In-the-wild robot teaching without in-the-wild robots," in *RSS*, 2024.

[17] Z. Zhaxizhuoma, K. Liu, C. Guan, Z. Jia, Z. Wu, X. Liu, T. Wang, S. Liang *et al.*, "Fast-UMI: A scalable and hardware-independent universal manipulation interface with dataset," in *CoRL*, 2025.

[18] Y. Dou, F. Yang, Y. Liu, A. Loquercio, and A. Owens, "Tactile-augmented radiance fields," in *CVPR*, 2024.

[19] N. Cheng, J. Xu, C. Guan, J. Gao, W. Wang, Y. Li, F. Meng, J. Zhou, B. Fang *et al.*, "Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation," *Information Fusion*, 2025.

[20] F. Tao and M. Salmeron, "In situ studies of chemistry and structure of materials in reactive environments," *Science*, 2011.

[21] W. Zheng, P. Chai, J. Zhu, and K. Zhang, "High-resolution in situ structures of mammalian respiratory supercomplexes," *Nature*, 2024.

[22] S. Chen, C. Wang, K. Nguyen, L. Fei-Fei, and C. K. Liu, "ARCap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback," in *ICRA*, 2025.

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

[24] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang, "Bunny-VisionPro: Real-time bimanual dexterous teleoperation for imitation learning," *arXiv preprint arXiv:2407.03162*, 2024.

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

[26] T. Buamanee, M. Kobayashi, Y. Uranishi, and H. Takemura, "Bi-ACT: Bilateral control-based imitation learning via action chunking with transformer," in *AIM*, 2024.

[27] F. Liu, C. Li, Y. Qin, A. Shaw, J. Xu, P. Abbeel, and R. Chen, "ViTaMIn: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface," *arXiv preprint arXiv:2504.06156*, 2025.

[28] Y. Pan, R. Qiao, L. Chen, K. Chitta, L. Pan, H. Mai, Q. Bu, H. Zhao, C. Zheng, P. Luo, and H. Li, "Agility Meets Stability: Versatile humanoid control with heterogeneous data," in *ICRA*, 2025.

[29] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *RSS*, 2023.

[30] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *IJRR*, 2024.

[31] A. Brygo, I. Sarakoglou, N. Garcia-Hernandez, and N. Tsagarakis, "Humanoid robot teleoperation with vibrotactile based balancing feedback," in *EuroHaptics*, 2014.

[32] H.-J. Huang, M. Kaess, and W. Yuan, "NormalFlow: Fast, robust, and accurate contact-based object 6dof pose tracking with vision-based tactile sensors," *RAL*, 2025.

[33] J. Kerr, H. Huang, A. Wilcox, R. Hoque, J. Ichnowski, R. Calandra, and K. Goldberg, "Self-supervised visuo-tactile pretraining to locate and follow garment features," in *RSS*, 2023.

[34] L. Fu *et al.*, "A touch, vision, and language dataset for multimodal alignment," *arXiv preprint arXiv:2402.13232*, 2024.

[35] R. Gao, Y. Dou, H. Li, T. Agarwal, J. Bohg, Y. Li, L. Fei-Fei, and J. Wu, "The objectfolder benchmark: Multisensory learning with neural and real objects," in *CVPR*, 2023.

[36] T. Li, Y. Yan, C. Yu, J. An, Y. Wang, X. Zhu, and G. Chen, "Vtg: A visual-tactile dataset for three-finger grasp," *RAL*, 2024.

[37] Q. Liu, Y. Cui, Z. Sun, G. Li, J. Chen, and Q. Ye, "VTDexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning," in *ICLR*, 2025.

[38] X. Zhu, B. Huang, and Y. Li, "Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper," in *NeurIPS*, 2025.

[39] J. Ren, J. Zou, and G. Gu, "MC-Tac: Modular camera-based tactile sensor for robot gripper," in *ICIRA*, 2023.

[40] A. C. Zoeller and K. Drewing, "A systematic comparison of perceptual performance in softness discrimination with different fingers," *Attention, Perception, & Psychophysics*, 2020.

[41] P. Manceron, "IKPy," https://github.com/Phylliade/ikpy, 2016.

[42] A. George, S. Gano, P. Katragadda, and A. B. Farimani, "Visuo-tactile pretraining for cable plugging," in *ICRA*, 2024.

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.