

ManipForce: Force-Guided Policy Learning with Frequency-Aware Representation for Contact-Rich Manipulation

Geonhyup Lee¹, Yeongjin Lee¹, Kangmin Kim¹, Seongju Lee¹, Sangjun Noh¹, Seunghyeok Back², Kyoobin Lee^{1†}

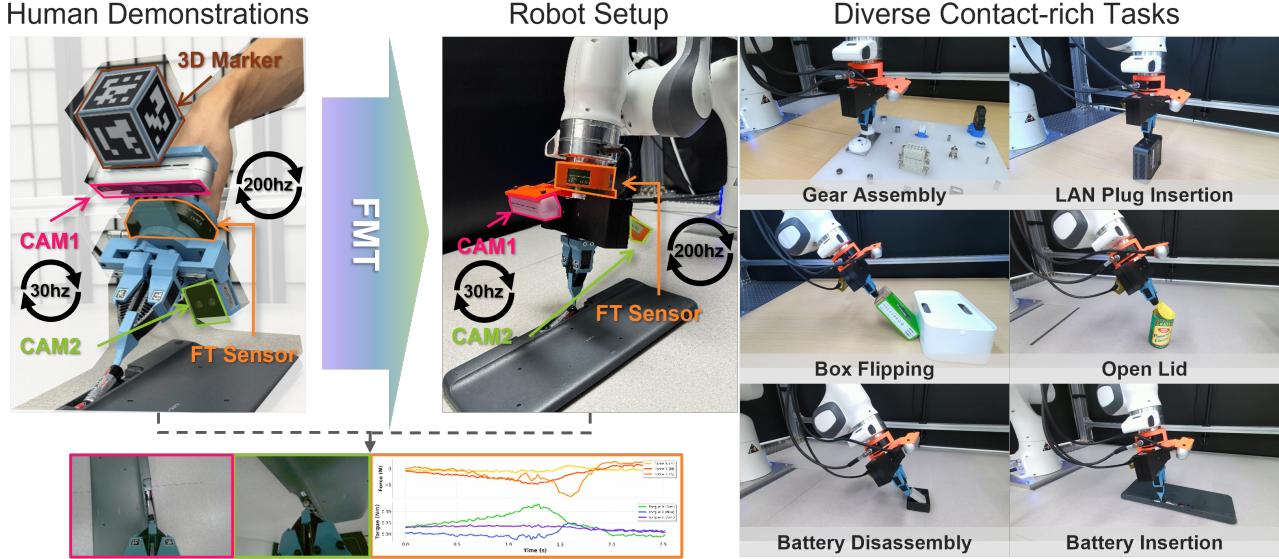


Fig. 1: Overview of our framework. (Left) **ManipForce**: a handheld system with dual cameras and a wrist-mounted F/T sensor capturing RGB–F/T data from human demonstrations. (Middle) The same configuration on a robot enables direct transfer to real-world execution. (Right) Six contact-rich tasks—gear assembly, LAN plug insertion, box flipping, open lid, battery disassembly, and battery insertion—used to evaluate our **FMT** model.

Abstract—Contact-rich manipulation tasks such as precision assembly require precise control of interaction forces, yet existing imitation learning methods rely mainly on vision-only demonstrations. We propose **ManipForce**, a handheld system designed to capture high-frequency force-torque (F/T) and RGB data during natural human demonstrations for contact-rich manipulation. Building on these demonstrations, we introduce the Frequency-Aware Multimodal Transformer (FMT). FMT encodes asynchronous RGB and F/T signals using frequency- and modality-aware embeddings and fuses them via bi-directional cross-attention within a transformer diffusion policy. Through extensive experiments on six real-world contact-rich manipulation tasks—such as gear assembly, box flipping, and battery insertion—FMT trained on **ManipForce** demonstrations achieves robust performance with an average success rate of 83% across all tasks, substantially outperforming RGB-only baselines. Ablation and sampling-frequency analyses further confirm that incorporating high-frequency F/T data and cross-modal integration improves policy performance, especially in tasks demanding high precision and stable contact. Hardware, software, and video demos are available at: <https://sites.google.com/view/manipforce/>.

¹ G. Lee, Y. Lee, K. Kim, S. Lee, S. Noh, and K. Lee are with the Department of AI Convergence, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea.

² S. Back is with the Department of AI Machinery, Korea Institute of Machinery & Materials (KIMM), Daejeon 34103, Republic of Korea.

† Corresponding author: Kyoobin Lee kyoobinlee@gist.ac.kr

I. INTRODUCTION

Contact-rich manipulation tasks such as precise assembly [1]–[4], battery disassembly [5], and non-prehensile handling [6] require high precision and **force-aware manipulation**. Humans naturally perceive contact forces and their subtle changes when assembling parts, adjusting their strategies accordingly. Yet most robotic approaches rely solely on visual demonstrations, missing the rich F/T information humans provide.

Recent advances in imitation learning [7]–[9] have demonstrated strong potential for dexterous and contact-rich manipulation by learning directly from human demonstrations. However, these methods still rely on high-quality demonstration data, which is costly and difficult to collect for fine-grained interactions. Hand-held data collection systems such as UMI [10] have been proposed to address this challenge by enabling natural human demonstrations without the expertise requirements and remote-control limitations of teleoperation. While effective for simplifying demonstration collection, UMI does not capture force-torque (F/T) information, which is essential for accurately modeling contact behaviors. More recent work [11] combines visual and F/T data but relies on point clouds to represent the scene, which introduces

complex setup requirements and fundamentally limits the ability to perceive small objects and fine clearances essential for contact-rich manipulation. Furthermore, from a learning perspective, this approach down-samples high-frequency F/T signals to match the image frame rate, losing rich temporal information necessary for modeling contact dynamics.

To address these limitations, we introduce **ManipForce** a handheld system for simultaneous RGB–F/T data collection during natural human demonstrations, and the **Frequency-Aware Multimodal Transformer (FMT)**, which learns robust policies from the collected data for diverse, precise, and contact-rich manipulation tasks.

ManipForce consists of a dual handheld camera setup with a wrist-mounted F/T sensor to capture both visual and high-frequency force signals during human-guided demonstrations. This configuration enables robust perception of small objects, tight clearances, and fine-grained contacts, allowing collected demonstrations to transfer directly to robotic execution. We replace SLAM-based wrist tracking with 3D ArUco marker pose estimation to maintain accuracy during close-contact interactions without environmental dependencies, and apply tool gravity compensation to ensure precise and interaction-focused F/T measurements. We propose the **FMT**, which learns from asynchronous RGB (30 Hz) and F/T (>200 Hz) signals using a Transformer-based Diffusion Policy [7] architecture. To exploit the higher-frequency force signals relative to images, the model tokenizes both RGB and F/T inputs using learnable frequency and modality embeddings. This design enables the model to effectively handle heterogeneous modalities with asynchronous sampling rates. In addition, bi-directional cross-attention modules fuse complementary information across modalities. We evaluate our approach on six contact-rich manipulation tasks spanning precision assembly, non-prehensile manipulation, and complex disassembly, and observe significant performance gains over RGB-only baselines. Ablation studies further confirm that high-frequency F/T sensing, unified positional embeddings, and bi-directional cross-attention each make complementary contributions to robust multimodal policy learning.

Our main contributions are:

- We introduce **ManipForce**, a handheld RGB–F/T data collection system enabling diverse and fine-grained contact-rich manipulation demonstrations.
- We propose **FMT**, which handles inputs with asynchronous sampling rates through frequency-aware multimodal representation learning and cross-attention within a Transformer architecture, enabling robust policy learning for contact-rich manipulation.
- We demonstrate robust performance on diverse contact-rich manipulation tasks—including gear assembly, plug insertion, battery disassembly, and lid operations—consistently outperforming RGB-only baselines.

II. RELATED WORK

A. Human Demonstration Data Collection

Teleoperation-based human demonstration collection includes VR-based systems [12]–[14], 3D spacetomouse controllers [15], [16], and leader–follower systems with joint mapping [9], [17]–[22]. While leader–follower systems provide intuitive control, teleoperation takes more time, needs skilled operators, and lacks clear visual and haptic feedback. Hand-held data collection systems, such as UMI [10], have shown a promising direction, enabling intuitive data collection by allowing users to directly manipulate a hand-held gripper. This paradigm has since expanded to include additional sensing modalities—audio [23], tactile [24]–[27], and F/T [11], [28], [29]. However, current F/T data collection approaches remain limited. Specialized fingertip force sensors and fixed-wrist setups are required [29]. Others [11] combine visual and F/T data but rely on manual cropping for data preparation, depth-based point clouds for scene representation, and visual-SLAM tracking for pose estimation. These steps introduce complex setup requirements, make it difficult to capture small features or precise depth in tasks such as LAN plug insertion, and reduce accuracy during close-contact interactions. Our proposed system addresses these limitations by enabling direct RGB–F/T data collection with a hand-eye camera setup, supporting diverse and fine-grained manipulation tasks without task-specific calibration, and employing marker-based pose tracking to maintain accuracy during close-contact interactions.

B. Multimodal Imitation Learning

Recent advances in visual imitation learning [7], [8] have shown robustness by modeling complex action distributions directly from images. However, vision-only approaches omit F/T cues that are critical for contact-rich manipulation. To address this limitation, multimodal approaches combine vision with force sensing through various strategies, including feature-level concatenation [11], [21], [30], and dynamic modality weighting via contact prediction [31]. Other studies introduce cross-attention modules for vision–F/T integration [5], or adaptive compliance control with learned dynamic gains [6]. Parallel work on vision–tactile integration tackles similar challenges through feature concatenation [32], [33], force-guided cross-attention [34], and CLIP-based visual–tactile representation learning [24]. A key challenge across these multimodal settings is managing asynchronous, rate-mismatched inputs, where visual sensors operate at low frequencies while force or tactile sensors operate at much higher rates. Although recent efforts such as Reactive Diffusion Policy [35] introduce dual-architecture designs to address this issue, they still impose complex structures and restrict cross-modal interaction to low-frequency visual streams. We address these limitations with FMT, a Transformer-based Diffusion Policy that uses learnable modality and frequency embeddings together with bi-directional cross-modal attention, enabling

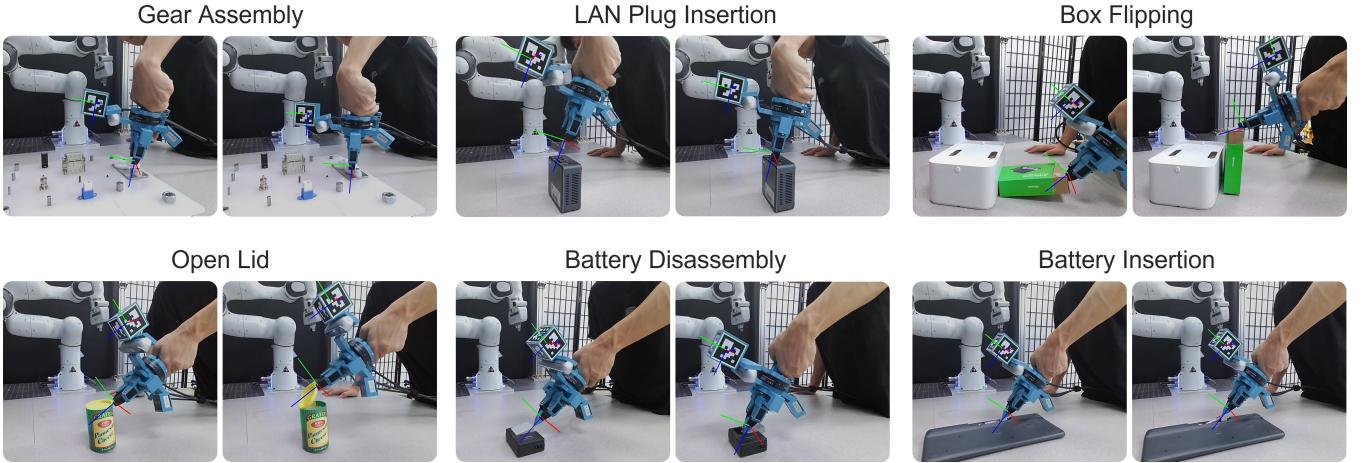


Fig. 2: Six contact-rich manipulation tasks used in our evaluation: Gear Assembly, LAN Plug Insertion, Box Flipping, Open Lid, Battery Disassembly, and Battery Insertion. These tasks require precise force control and multimodal feedback, testing our model’s ability to integrate high-frequency force signals with visual input for robust manipulation.

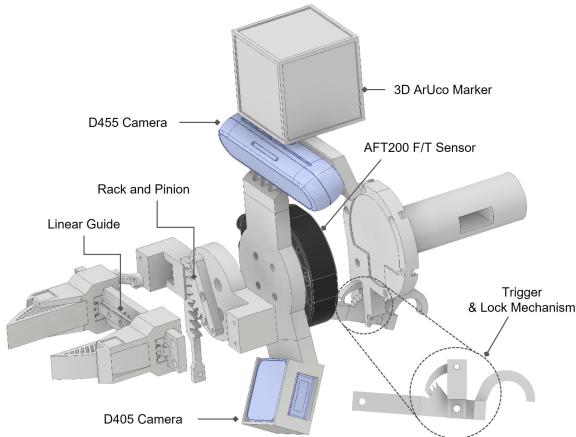


Fig. 3: Structure of the ManipForce. It includes dual RGB cameras, a wrist-mounted F/T sensor, and ArUco marker tracking. A rack-and-pinion gripper with trigger-lock and linear guides enables precise, stable human demonstrations.

frequency-aware multimodal representation learning across asynchronous multimodal inputs.

III. METHOD

In this section, we present our two-component framework combining data collection and policy learning. **ManipForce** (Section III-A) is a handheld RGB–F/T system that captures high-precision poses and high-bandwidth force signals during natural human demonstrations, yielding diverse and fine-grained data. Building on this data, the **FMT** (Section III-B) uses frequency-aware multimodal embeddings and bi-directional cross-attention to learn robust policies from asynchronous visual and force inputs.

A. ManipForce

System Design. Our gripper mechanism is purpose-built to make human-guided data collection intuitive, precise, and

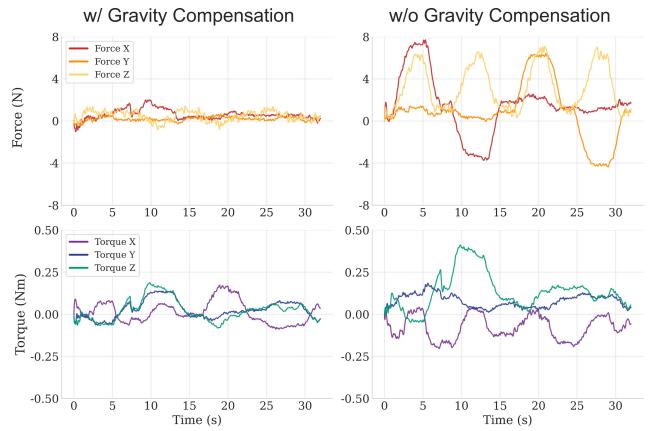


Fig. 4: Gravity compensation result in ManipForce. Tilting the system by $\pm 90^\circ$ about the X and Y axes shows that the measured F/T signals remain stable despite changes in orientation.

stable. As shown in Fig. 3, the design integrates several complementary features that improve usability and data quality. At its core, a rack-and-pinion mechanism with a familiar trigger interface drives parallel jaw motion, allowing users to perform natural and accurate grasping actions. A deformable pin-ray structure at the fingertips increases compliance and contact sensitivity, while integrated linear guides at the gripper attachment points suppress mechanical vibration and promote smoother, more precise manipulation. A trigger-lock mechanism holds the jaws closed after grasping, reducing user fatigue and ensuring that recorded signals primarily reflect task-relevant interaction forces. Finally, all gripper components are mounted downstream of the F/T sensor to capture the complete interaction forces during operation. All hardware and software components are released as open source at our project website.

RGB and Wrist-Pose Data Acquisition. The system

employs two hand-eye cameras (Intel RealSense D455 and D405) positioned to minimize occlusions and ensure clear visibility during precision manipulation tasks. Conventional SLAM-based wrist-pose estimation degrades near surfaces and cannot meet the precision required for fine assembly tasks such as gear or LAN plug insertion. To address this, we use 3D ArUco markers detected by an externally positioned Azure Kinect sensor, achieving robust, viewpoint-independent tracking with sub-millimeter accuracy [36]. End-effector actions are computed as 6-DOF pose deltas between consecutive frames and transformed from the marker reference frame to the robot’s TCP frame using CAD-derived transformation matrices. Additional ArUco markers attached to both gripper jaws provide real-time tracking of jaw positions. The velocity of these markers between consecutive frames is further analyzed to infer gripper open/close states, enabling detailed annotation of manipulation events. All RGB and wrist-pose data are recorded at 30 Hz.

F/T Data Acquisition. An AIDIN AFT200 F/T sensor mounted on the wrist records force/torque signals at 200 Hz. Because raw F/T measurements include both interaction forces and the weight of the tool, all F/T signals are gravity-compensated using IMU data from the D455 camera to isolate true contact forces and ensure consistency between the ManipForce system and robot platforms. The gravity vector from the IMU frame is transformed into the F/T sensor frame as

$$\mathbf{g}_{ft} = R_{imu}^{ft} \mathbf{g}_{imu},$$

where R_{imu}^{ft} denotes the rotation from the IMU to the F/T sensor frame. The gravitational wrench is then computed as

$$\mathbf{F}_g = m_{tool} \mathbf{g}_{ft}, \quad \boldsymbol{\tau}_g = \mathbf{r}_{com} \times \mathbf{F}_g,$$

and subtracted from the measured wrench to obtain the compensated values:

$$\mathbf{F}_{comp} = \mathbf{F}_{measure} - \mathbf{F}_g, \quad \boldsymbol{\tau}_{comp} = \boldsymbol{\tau}_{measure} - \boldsymbol{\tau}_g.$$

This gravity compensation procedure is applied identically during both handheld data collection and robot execution to ensure consistent F/T measurements across systems. To validate the effectiveness of this compensation, the handheld system was rotated about the x- and y-axes by $\pm 90^\circ$ while measuring the F/T signals (Fig. 4). Without compensation, force readings reached up to $\pm 8\text{N}$ and torques up to $0.4\text{N}\cdot\text{m}$, whereas with compensation, the residual forces were reduced to within $\pm 1\text{N}$ and torques to within $0.15\text{ N}\cdot\text{m}$, confirming that the procedure effectively removes gravity-induced bias.

B. Frequency-aware Multimodal Transformer

In this section, we present **FMT**, a model that learns frequency-aware multimodal representations to fuse low-rate RGB with high-frequency F/T data for contact-rich manipulation. Extending Diffusion Policy [7], FMT applies frequency-aware multimodal embeddings and bi-directional cross-attention to align visual cues with fine-grained force information. We outline its four components—multimodal tokenization, frequency-modality embeddings, cross-attention

fusion, and the diffusion-based policy head—that together enable robust learning from asynchronous multimodal inputs. The overall architecture of FMT is illustrated in Fig. 5.

RGB-Force Tokenization. Visual observations from dual hand-eye cameras are preprocessed through squared padding and resizing 256×256 resolution before being encoded using DINOv2-B [37] to extract visual representations, while F/T representation is extracted via a 1D CNN encoder to capture force dynamics. Specifically, the visual representations from the two hand-eye cameras are denoted by $\mathbf{T}_{cam1}, \mathbf{T}_{cam2} \in \mathbb{R}^{T_{img} \times L \times d}$, and the F/T representation by $\mathbf{T}_{ft} \in \mathbb{R}^{T_{ft} \times d}$. Here, T_{img} is the time horizon of the hand-eye camera, L is the number of visual tokens within an image, d is the model dimension, and T_{ft} is the time horizon of the F/T observations. We set $T_{img} = 2$, $T_{ft} = 8$, $L = 256$, and $d = 768$. To handle asynchronous inputs, we employ timestamp-based windowing where each visual frame captured at 30 Hz is aligned with corresponding F/T samples captured at 200+ Hz within the temporal boundaries defined by consecutive image frames. This preserves high-frequency force dynamics while maintaining correspondence with visual observations.

Frequency-aware Multimodal Embedding. To enable unified processing, we align the encoded modalities using three types of learnable positional embeddings: (i) spatial embeddings, $\mathbf{E}_{spatial} \in \mathbb{R}^{L \times d}$, which encode within-map positions for visual tokens; (ii) frequency-aware embeddings, $\mathbf{E}_{freq} \in \mathbb{R}^{T_{ft} \times d}$, which capture timestamp relationships across modalities with different sampling rates; and (iii) modality embeddings, $\mathbf{E}_{cam1}, \mathbf{E}_{cam2}, \mathbf{E}_{ft} \in \mathbb{R}^{1 \times d}$ that encode the source modality of each token. For the hand-eye cameras, frequency-aware embeddings are obtained by resampling \mathbf{E}_{freq} to $\mathbb{R}^{T_{img} \times d}$ via linear interpolation *to approximately align with the cameras’ asynchronous timestamps*. These embeddings are added to the visual and F/T tokens and are learned during training, enabling the transformer to model the spatial, frequency, and modality-specific context of each token in the heterogeneous multimodal sequence. As a result, the unified visual tokens $\mathbf{T}'_{cam} \in \mathbb{R}^{2LT_{img} \times d}$ and the F/T tokens $\mathbf{T}'_{ft} \in \mathbb{R}^{T_{ft} \times d}$ are used as inputs to the bi-directional cross-attention module.

Bi-directional Cross-Attention. The architecture employs bi-directional cross-attention mechanisms to facilitate information exchange between modalities. Visual tokens attend to F/T tokens to incorporate contact dynamics, while F/T tokens attend to visual features to understand spatial context. This design enables learning of rich cross-modal representations that capture interdependencies between visual scenes and physical interactions. Mathematically, the enhanced visual and F/T tokens can be represented as follows:

$$\mathbf{T}''_{img} = CA_{img \leftarrow ft}(\mathbf{Q} = \mathbf{T}'_{img}, \mathbf{K} = \mathbf{T}'_{ft}, \mathbf{V} = \mathbf{T}'_{ft}) \quad (1)$$

$$\mathbf{T}''_{ft} = CA_{ft \leftarrow img}(\mathbf{Q} = \mathbf{T}'_{ft}, \mathbf{K} = \mathbf{T}'_{img}, \mathbf{V} = \mathbf{T}'_{img}), \quad (2)$$

where $CA_{a \leftarrow b}$ denotes cross-attention that uses queries from a and keys/values from b (i.e., $\mathbf{Q} = \mathbf{X}_a$, $\mathbf{K} = \mathbf{V} = \mathbf{X}_b$). Finally, we obtain the unified observation representation by concatenating the enhanced visual and F/T tokens, applying

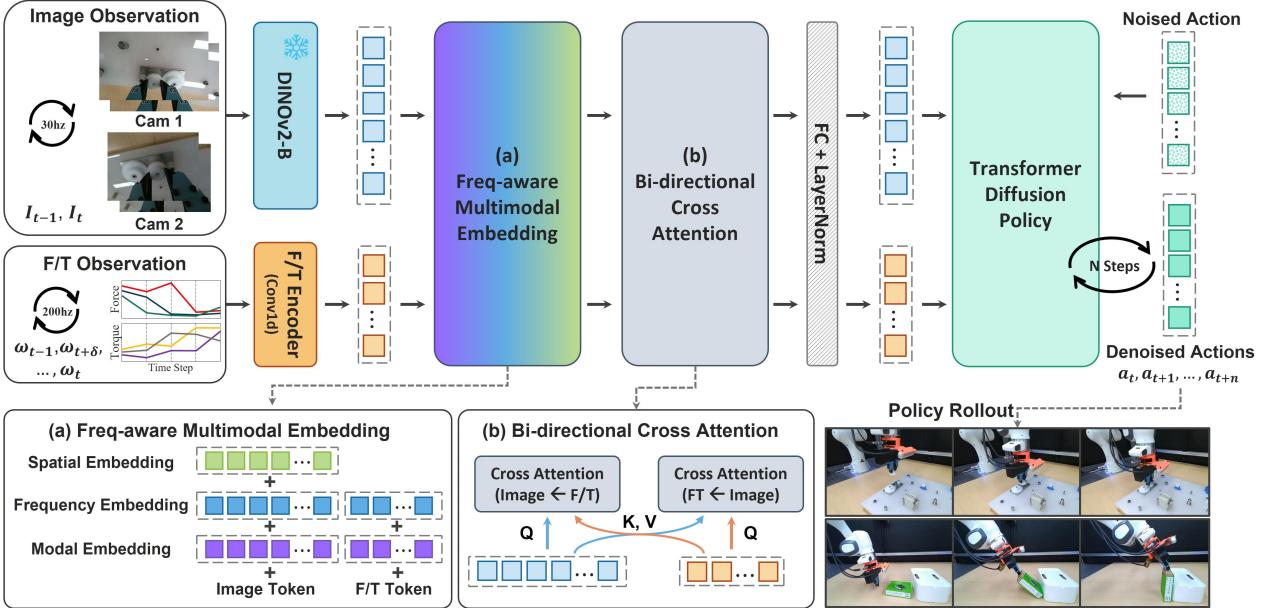


Fig. 5: Architecture of FMT. Asynchronous RGB (30 Hz) and F/T (200 Hz) signals are encoded with frequency-aware multimodal embeddings, fused via bi-directional cross-attention, and passed through a Transformer Diffusion Policy to produce robust actions for precise, contact-rich manipulation tasks.

a fully connected layer, and then layer normalization:

$$\mathbf{T}_{\text{obs}} = \text{LN}(\text{FC}(\text{Cat}(\mathbf{T}_{\text{img}}'', \mathbf{T}_{\text{ft}}''))) \in \mathbb{R}^{(2LT_{\text{img}} + T_{\text{ft}}) \times d}. \quad (3)$$

Transformer Diffusion Policy. Following the Time-series Diffusion Transformer [7] architecture, the unified multimodal tokens \mathbf{T}_{obs} serve as conditioning input for action generation. At each denoising step, noisy action embeddings attend to the multimodal observation features through cross-attention, while maintaining causal self-attention over previous action tokens. The noise prediction network generates action sequences through iterative denoising steps, naturally handling variable-length sequences from different sampling rates without requiring explicit synchronization strategies.

IV. EXPERIMENTAL RESULTS

We evaluate the proposed **FMT** on six real-world contact-rich manipulation tasks to validate its ability to fuse high-frequency force and RGB data. Section IV-B compares FMT with an RGB-only baseline to quantify the benefits of incorporating high-frequency force information. Section IV-C isolates the effects of F/T sensing, frequency-aware multimodal embeddings, and cross-attention via ablation. Section IV-D examines how varying the F/T sampling rate from 30–200 Hz affects task performance. Finally, Section IV-E analyzes data-collection efficiency, comparing ManipForce with direct human demonstrations and teleoperation system.

A. Experimental Setup

1) *Evaluation Tasks:* We evaluate our approach on six contact-rich manipulation tasks that demand precise force control and multimodal feedback (Fig. 2). The tasks include **gear assembly**, which demands precise rotational alignment

with controlled axial force, **LAN plug insertion**, which requires sub-millimeter accuracy and carefully regulated insertion forces to avoid damaging component, and **box flipping**, a non-prehensile manipulation task requiring coordinated push-and-roll motions guided by contact sensing. We also consider **open lid**, where the gripper must detect and apply force on a small handle to lift it; **battery disassembly**, where a tool is inserted into narrow gaps to sense contact force and lift batteries safely; and **battery insertion**, which involves inserting a spring-loaded battery while maintaining the correct force and direction to prevent slipping or ejection. Together, these tasks evaluate our model’s ability to integrate high-frequency force signals with visual feedback for accurate and robust manipulation. For all tasks, approximately 100 demonstration episodes were collected for training, and each task was evaluated over 20 trials with randomized initial robot and object poses.

2) *Robot Setup:* We evaluate our approach on a 7-DOF Franka Panda robot equipped with an AIDIN AFT200 F/T sensor mounted at the wrist, identical to the sensor used in the ManipForce data collection system. Tool gravity compensation is applied to the robot’s F/T measurements to ensure consistency with the handheld data collection system. A compliance controller is employed to improve contact safety and stability during manipulation. The hand-eye cameras are positioned to match the TCP-to-lens distances used in ManipForce, preserving geometric consistency between demonstration collection and robot execution (see Fig. 1).

B. Does High-Frequency Force-Aware Policy Learning Improve Performance on Contact-Rich Manipulation?

We compare two models based on the Transformer Diffusion Policy architecture. The **RGB-only** model [7] replaces

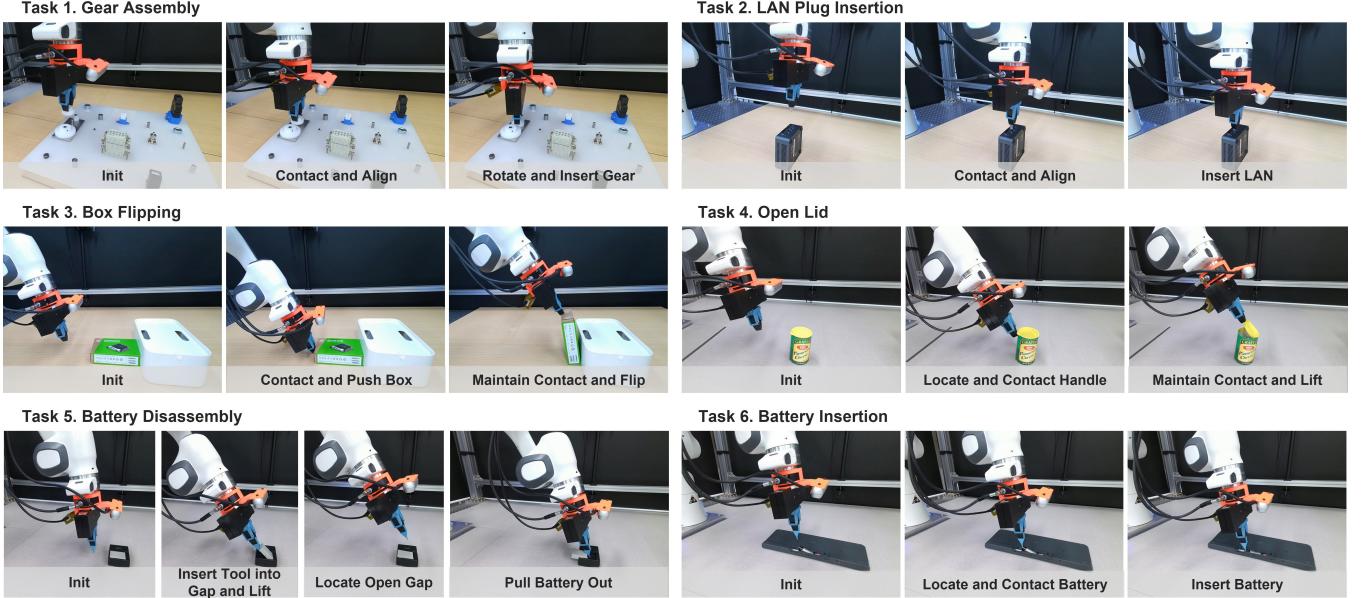


Fig. 6: Policy rollout examples from the proposed FMT across the six contact-rich manipulation tasks used for evaluation. Each sequence shows the trained policy executing key stages of each task using integrated RGB–F/T feedback.

TABLE I: Success rates on contact-rich manipulation tasks comparing our RGB–F/T model with an RGB-only baseline. Each entry reports the proportion of successful trials over 20 evaluation episodes per task.

Method	Gear Assembly	LAN Plug Insertion	Box Flipping	Open Lid	Battery Disassembly	Battery Insertion
RGB-only [7]	0.35	0.40	0.05	0.20	0.20	0.10
FMT	0.95	0.85	0.90	1.00	0.65	0.60

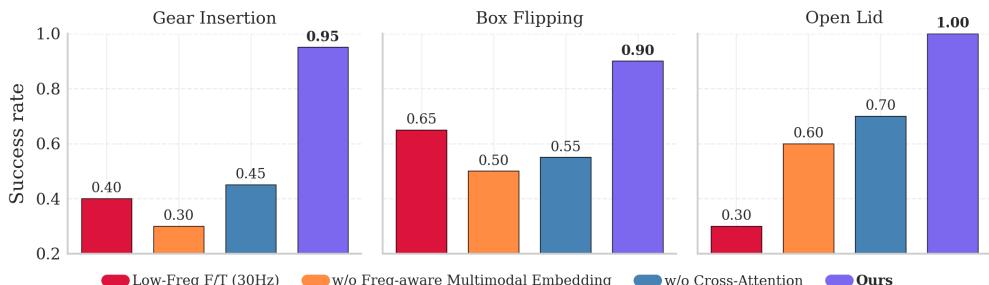


Fig. 7: Ablation study on three representative tasks (Gear Assembly, Box Flipping, and Open Lid). We compare the full FMT model with low-frequency F/T input (30 Hz), without positional embeddings, and without cross-attention. FMT consistently outperforms all ablated variants, highlighting the importance of high-frequency F/T sensing, unified positional embeddings, and bi-directional cross-attention for robust manipulation performance.

the original ResNet-18 image encoder with DINOv2-B and uses only RGB tokens as input to the Transformer Diffusion Policy. In contrast, our proposed **FMT** extends the same backbone by incorporating F/T inputs and employing frequency-aware multimodal embeddings with bi-directional cross-attention, enabling fusion of visual and force information at their native sampling rates (Section III-B). As shown in Table I, FMT achieves higher success rates across all six contact-rich tasks, especially in scenarios requiring precise force control or tight clearances. On average, it reaches

83% success compared to 22% for the RGB-only model, highlighting the benefit of high-frequency force-aware policy learning. For example, in **gear and battery assembly**, which requires repeated fine alignment under rapidly changing forces, FMT maintains precise insertion and stable contact far beyond the RGB-only baseline. In **box flipping**, where sustained contact and continuous force regulation are essential, FMT achieves smooth, reliable manipulation. Similarly, in **open lid**, which hinges on capturing a single transient high-frequency force spike, FMT detects and exploits this

cue, achieving robust task execution where the baseline fails. Overall, FMT combines stable contact control with transient event detection, yielding robust and reliable performance across diverse manipulation scenarios (see Fig. 6)

C. How Do FMT Modules Influence Performance Across Contact-Rich Tasks?

We conducted an ablation study to assess the contribution of FMT’s three core modules: high-frequency force sensing, modal-frequency embeddings, and bi-directional cross-attention. We compared the full FMT against three ablated variants: **Low-Freq F/T (30 Hz)** to evaluate the impact of downsampling the force stream; **w/o Freq-aware Multi-modal Embeddings** to test the importance of aligning asynchronous RGB and force inputs; and **w/o Cross-Attention** to measure the effect of disabling dynamic vision–force fusion. Fig. 7 shows the success rates of all variants across representative tasks.

Across all six contact-rich tasks, the full FMT achieved the highest success rates, with particularly large gains in tasks demanding fine force modulation or narrow clearances. In **gear assembly**, which requires repeated fine alignment under rapidly changing forces, removing the modal-frequency embeddings caused the steepest performance drop, underscoring their role in precise frequency and cross-modal alignment. **Box flipping**, dominated by slower quasi-static forces, was less sensitive to high-frequency loss but still benefited from both embeddings and cross-attention, showing that even low-frequency tasks gain from improved alignment and fusion. By contrast, **open lid**, which hinges on a single transient high-frequency force spike, exhibited the sharpest degradation under low-frequency input—confirming the need to preserve high-frequency force signatures for critical event detection.

Overall, these results highlight the complementary roles of FMT’s modules. High-frequency sensing preserves fine-grained force signatures, modal-frequency embeddings synchronize heterogeneous streams for coherent integration, and bi-directional cross-attention adaptively fuses vision and force for stable, robust policies. Together, these elements form the frequency-aware multimodal representations that drive FMT’s strong and consistent performance across diverse contact-rich manipulation tasks.

D. How Does F/T Sampling Frequency Affect Task Performance?

Fig. 8 presents the success rate of the **gear assembly** task at different F/T sampling frequencies (30 Hz, 60 Hz, 120 Hz, and 200 Hz). Performance rises monotonically from 0.40 at 30 Hz to 0.95 at 200 Hz, confirming that our model benefits from higher-frequency F/T inputs. At lower sampling rates, the model receives a coarser frequency representation of contact forces, making it harder to detect transient sticking or jamming events and to perform fine rotational adjustments after insertion. In contrast, higher-frequency sensing captures short-duration force spikes and

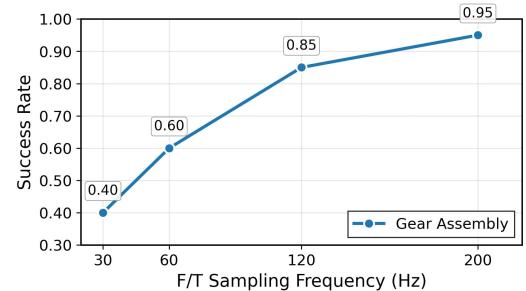


Fig. 8: Gear Assembly success rates across F/T sampling frequencies, indicating improved performance at higher frequencies.

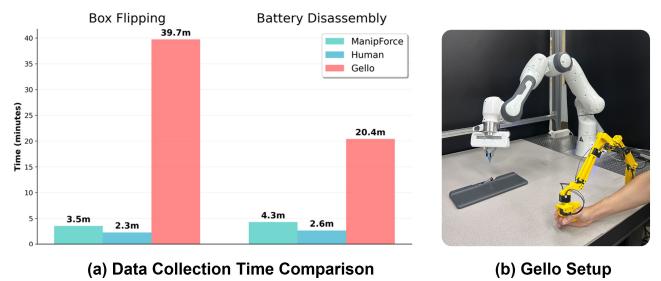


Fig. 9: Data collection efficiency comparison among human demonstrations, our ManipForce, and the teleoperation-based Gello, showing ManipForce achieves near-human speed while teleoperation is slower due to the difficulty of maintaining contact and precise force control.

subtle torque changes, enabling more precise corrective actions and smoother alignment. These results demonstrate that high-frequency multimodal sensing directly enhances policy performance in contact-rich assembly tasks and validate our model’s ability to exploit high-bandwidth F/T data.

E. How Effectively Does ManipForce Capture Contact-Rich Demonstrations?

We compare data collection efficiency across three settings—**human demonstrations**, our **ManipForce** handheld system, and the teleoperation-based **Gello** system [17] (shown in Fig. 9b). To quantify efficiency, we measured the time required to record demonstrations for two representative tasks—**box flipping** and **battery disassembly**—across these settings (results in Fig. 9a). ManipForce achieves efficiency comparable to direct human demonstrations, while Gello requires dramatically longer recording times. Beyond efficiency, teleoperation frameworks such as Gello inherently lack haptic feedback, making it difficult for operators to apply accurate interaction forces during contact-rich tasks. In contrast, ManipForce leverages natural human operation through a handheld system, ensuring that demonstrations capture both task-relevant motions and the precise interaction forces needed for successful execution.

V. CONCLUSION

We presented ManipForce, a handheld system with dual hand-eye cameras and a wrist-mounted F/T sensor that captures high-frequency multimodal data during natural human demonstrations, enabling direct transfer to robot execution without task-specific calibration. Building on these demonstrations, we introduced FMT, which integrates high-frequency F/T and visual signals for policy learning in contact-rich manipulation, effectively handling their asynchronous sampling rates to achieve precise and stable execution. Across six diverse manipulation tasks, policies trained with FMT achieved an average success rate of about 83%, substantially outperforming RGB-only baselines and confirming the benefit of high-bandwidth sensing and cross-modal fusion for robust policy learning. While our current study demonstrates strong performance on diverse contact-rich tasks, we plan to extend our framework to longer-horizon and more dexterous tasks requiring more sophisticated gripper behaviors. These directions aim to further enhance the generality and scalability of ManipForce for complex real-world manipulation scenarios.

ACKNOWLEDGMENTS

This work was supported by the Technology Innovation Program (RS-2024-00442029, Development of Tactile Intelligence in Robotic Hands Based on Tactile Data Learning to Manipulate Irregular Multiple Types of Objects and RS-2024-00423940, Development of Humanoid Robots That Feel Like Humans, Communicate, and Grow through Learning) funded by the Ministry of Trade Industry & Energy(MOTIE, Korea).

REFERENCES

- [1] J. Luo, E. Solowjow, C. Wen, J. A. Ojea, A. M. Agogino, A. Tamar, and P. Abbeel, “Reinforcement learning on variable impedance controller for high-precision robotic assembly,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3080–3087.
- [2] Z. Zhang, Y. Wang, Z. Zhang, L. Wang, H. Huang, and Q. Cao, “A residual reinforcement learning method for robotic assembly using visual and force information,” *Journal of Manufacturing Systems*, vol. 72, pp. 245–262, 2024.
- [3] W. Chen, C. Zeng, H. Liang, F. Sun, and J. Zhang, “Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly,” *IEEE Transactions on Cybernetics*, 2023.
- [4] G. Lee, J. Lee, S. Noh, M. Ko, K. Kim, and K. Lee, “Polyfit: A peg-in-hole assembly framework for unseen polygon shapes via sim-to-real adaptation,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 533–540.
- [5] J. H. Kang, S. Joshi, R. Huang, and S. K. Gupta, “Robotic compliant object prying using diffusion policy guided by vision and force observations,” *IEEE Robotics and Automation Letters*, 2025.
- [6] 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,” *arXiv preprint arXiv:2410.09309*, 2024.
- [7] 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,” *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [8] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, “3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations,” *arXiv preprint arXiv:2403.03954*, 2024.
- [9] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware,” *arXiv preprint arXiv:2304.13705*, 2023.
- [10] 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,” *arXiv preprint arXiv:2402.10329*, 2024.
- [11] 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,” *arXiv preprint arXiv:2410.07554*, 2024.
- [12] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, “Open teach: A versatile teleoperation system for robotic manipulation,” *arXiv preprint arXiv:2403.07870*, 2024.
- [13] 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.
- [14] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, “Open-television: Teleoperation with immersive active visual feedback,” *arXiv preprint arXiv:2407.01512*, 2024.
- [15] J. Luo, C. Xu, F. Liu, L. Tan, Z. Lin, J. Wu, P. Abbeel, and S. Levine, “Fmb: a functional manipulation benchmark for generalizable robotic learning,” *The International Journal of Robotics Research*, vol. 44, no. 4, pp. 592–606, 2025.
- [16] L. Ankile, A. Simeonov, I. Shenfeld, and P. Agrawal, “Juicer: Data-efficient imitation learning for robotic assembly,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 5096–5103.
- [17] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, “Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 156–12 163.
- [18] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, “Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 031–15 038.
- [19] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, “Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation,” *arXiv preprint arXiv:2408.11805*, 2024.
- [20] J. Huang, K. Chen, J. Zhou, X. Lin, P. Abbeel, Q. Dou, and Y. Liu, “Dih-tele: Dexterous in-hand teleoperation framework for learning multiobjects manipulation with tactile sensing,” *IEEE/ASME Transactions on Mechatronics*, 2025.
- [21] J. J. Liu, Y. Li, K. Shaw, T. Tao, R. Salakhutdinov, and D. Pathak, “Factr: Force-attending curriculum training for contact-rich policy learning,” *arXiv preprint arXiv:2502.17432*, 2025.
- [22] N. Myers, O. Kwon, S. Yamsani, and J. Kim, “Child (controller for humanoid imitation and live demonstration): a whole-body humanoid teleoperation system,” *arXiv preprint arXiv:2508.00162*, 2025.
- [23] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, “Maniwav: Learning robot manipulation from in-the-wild audio-visual data,” *arXiv preprint arXiv:2406.19464*, 2024.
- [24] 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.
- [25] X. Zhu, B. Huang, and Y. Li, “Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper,” *arXiv preprint arXiv:2507.15062*, 2025.
- [26] L. Wu, C. Yu, J. Ren, L. Chen, R. Huang, G. Gu, and H. Li, “Freetacman: Robot-free visuo-tactile data collection system for contact-rich manipulation,” *arXiv preprint arXiv:2506.01941*, 2025.
- [27] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, “Tactile-vla: Unlocking vision-language-action model’s physical knowledge for tactile generalization,” *arXiv preprint arXiv:2507.09160*, 2025.
- [28] A. Adeniji, Z. Chen, V. Liu, V. Pattabiraman, R. Bhirangi, S. Haldar, P. Abbeel, and L. Pinto, “Feel the force: Contact-driven learning from humans,” *arXiv preprint arXiv:2506.01944*, 2025.
- [29] C. Chen, Z. Yu, H. Choi, M. Cutkosky, and J. Bohg, “Dexforce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation,” *IEEE Robotics and Automation Letters*, 2025.
- [30] T. Yang, Y. Jing, H. Wu, J. Xu, K. Sima, G. Chen, Q. Sima, and T. Kong, “Moma-force: Visual-force imitation for real-world mobile manipulation,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6847–6852.
- [31] Z. He, H. Fang, J. Chen, H.-S. Fang, and C. Lu, “Foar: Force-aware reactive policy for contact-rich robotic manipulation,” *arXiv preprint arXiv:2411.15753*, 2024.
- [32] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, “3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing,” *arXiv preprint arXiv:2410.24091*, 2024.

- [33] Q. K. Luu, P. Zhou, Z. Xu, Z. Zhang, Q. Qiu, and Y. She, "Manifeel: Benchmarking and understanding visuotactile manipulation policy learning," *arXiv preprint arXiv:2505.18472*, 2025.
- [34] J. Li, T. Wu, J. Zhang, Z. Chen, H. Jin, M. Wu, Y. Shen, Y. Yang, and H. Dong, "Adaptive visuo-tactile fusion with predictive force attention for dexterous manipulation," *arXiv preprint arXiv:2505.13982*, 2025.
- [35] 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," *arXiv preprint arXiv:2401.12349*, 2024.
- [36] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.
- [37] M. Oquab, T. Dariset, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.