FoAR: Force-Aware Reactive Policy for Contact-Rich Robotic Manipulation

Zihao He*, Hongjie Fang*, Jingjing Chen, Hao-Shu Fang† and Cewu Lu†
Shanghai Jiao Tong University

***Abstract*— Contact-rich tasks present significant challenges for robotic manipulation policies due to the complex dynamics of contact and the need for precise control. Vision-based policies often struggle with the skill required for such tasks, as they typically lack critical contact feedback modalities like force/torque information. To address this issue, we propose FoAR, a force-aware reactive policy that combines high-frequency force/torque sensing with visual inputs to enhance the performance in contact-rich manipulation. Built upon the RISE policy, FoAR incorporates a multimodal feature fusion mechanism guided by a future contact predictor, enabling dynamic adjustment of force/torque data usage between non-contact and contact phases. Its reactive control strategy also allows FoAR to accomplish contact-rich tasks accurately through simple position control. Experimental results demonstrate that FoAR significantly outperforms all baselines across various challenging contact-rich tasks while maintaining robust performance under unexpected dynamic disturbances. Project website: https://tonyfang.net/FoAR/.**

I. Introduction

Contact-rich manipulation is an essential field in robotics, involving tasks that require sustained, intricate contact with objects or environments [21]. Such tasks, including assembly [9, 26], wiping [11, 16], and peeling [2, 15], are inherently challenging due to the complex dynamics of force and precise control required. Unlike simple pick-and-place operations [27], contact-rich manipulation demands nuanced interaction and real-time adaptation to variations in object properties. As a result, developing effective algorithms and learning models for contact-rich manipulation is crucial for enabling more versatile and interactive robot systems.

In recent years, significant progress has been made in vision-based robotic manipulation policies [1, 3, 5, 13, 18, 23, 25, 28, 29]. However, these policies often fall short of achieving the dexterity required for contact-rich manipulations, as they typically lack crucial contact feedback, such as force/torque and tactile information.

In contact-rich manipulation, integrating force/torque sensing offers an intuitive and versatile approach by directly capturing the physical interactions between the robot and its environment since contact inherently produces forces and torques. In addition to the force and torque information, several previous works seek to improve the ability of the robot in contact-rich manipulation by incorporating other auxiliary modalities like audio [7, 14, 16, 17], and tactile [6, 7, 12, 14, 19]. While prior studies [11, 15, 24] have improved contact-rich task performance by incorporating the force/torque modality, they often *combine force/torque data with vision data through the whole manipulation process, ignoring the fact that force/torque are sparsely activated*. In practice, tasks like wiping involve multiple phases, such as picking up an eraser, performing the wiping, and placing the eraser down. Among these phases, only the wiping phase requires significant contact interactions. During non-contact phases of the task, the inherent noise in force/torque data from real-world sensors might degrade policy performance.

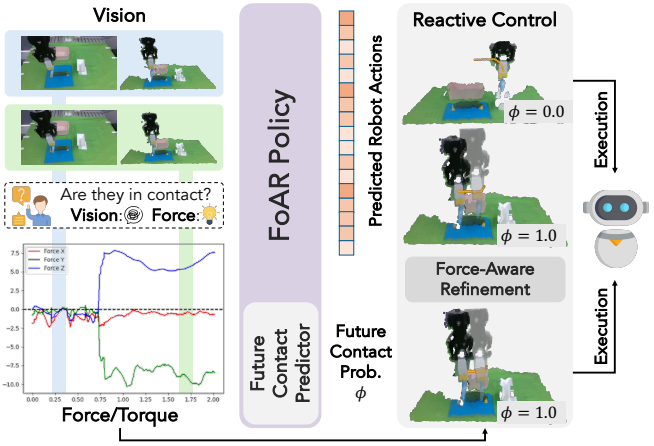


Fig. 1: Overview of the FoAR Policy for Contact-Rich Robotic Manipulations. Vision alone struggles to distinguish contact from non-contact states in contact-rich tasks, underscoring the need for integrating force/torque information. Our FoAR policy combines vision and force/torque inputs to predict robot actions along with a future contact probability ϕ. Reactive control then refines actions dynamically based on current and predicted future contact states, enabling precise, force-aware manipulations for contact-rich tasks.

This paper introduces FoAR, a force-aware reactive policy designed for contact-rich robotic manipulation tasks. Building on the state-of-the-art real-world robot imitation policy RISE [23], FoAR effectively integrates high-frequency force/torque sensing with visual inputs by dynamically balancing the usage of force/torque data. This enables precise handling of complex contact dynamics while maintaining strong performance in non-contact phases. The co-design of the FoAR policy and its reactive control strategy further enhances its contact-rich task performance through simple position control. With only 50 demonstrations per task, FoAR significantly outperforms baselines across various challenging contact-rich manipulation tasks.

* Equal Contribution.
† Hao-Shu Fang and Cewu Lu are the corresponding authors.
Emails: {he0610, galaxies, jjchen20}@sjtu.edu.cn, fhaoshu@gmail.com, lucewu@sjtu.edu.cn

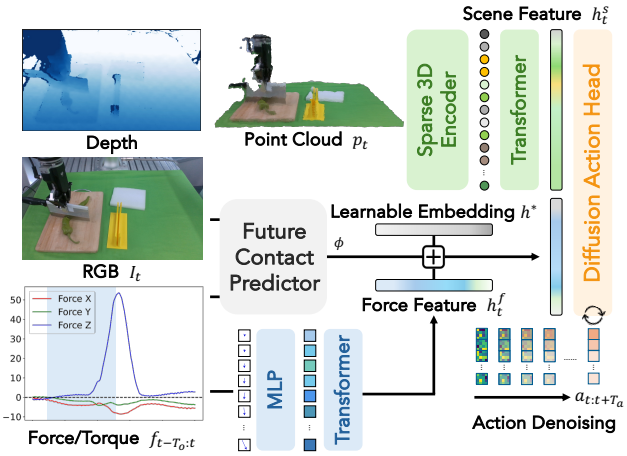


Fig. 2: FoAR Architecture. FoAR consists of a point cloud encoder [23], a force/torque encoder, a future contact predictor, and a diffusion action head [3]. The scene features and force features are fused under the guidance of the future contact predictor.

Algorithm 1 FoAR Inference with Reactive Control

1: ***buffer***.clear();
2: ***contact_buffer***.clear(); ▷ clear the temporal ensemble buffer.
3: **for** timestep $t \leftarrow 0$ to $N_{\text{max}} - 1$ **do**
4: **if** t mod $N_{\text{inference}} = 0$ **then** ▷ at the inference step.
5: $p_t, I_t, f_{t-T_o:t}, q_t \leftarrow$ ***agent***.perception(); ▷ perception.
6: $\phi, a_{t:t+T_a} \leftarrow \text{FoAR}(p_t, f_{t-T_o:t}, I_t)$; ▷ inference.
7: **if** $\phi < \delta_\phi$ **then** ▷ non-contact phase.
8: ***buffer***.add($a_{t:t+T_a}$);
9: **else** ▷ contact phase.
10: **if** force(f_t) $< \delta_f$ and torque(f_t) $< \delta_t$ **then**
11: ▷ insufficient force/torque detected.
12: $\mathbf{d} \leftarrow \text{avg}(a_{t:t+T_f}).\text{pos} - q_t.\text{pos}$;
13: $a_{t:t+T_a}.\text{pos} \leftarrow a_{t:t+T_a}.\text{pos} + \epsilon \cdot \mathbf{d}/\|\mathbf{d}\|_2$;
14: ▷ update actions towards predicted direction.
15: **end if**
16: ***contact_buffer***.add($a_{t:t+T_a}$);
17: **end if**
18: **end if**
19: $a_t \leftarrow$ ***buffer***.get(t) **if** $\phi < \delta_\phi$ **else** ***contact_buffer***.get(t);
20: ***agent***.execute(a_t); ▷ retrieve and execute the action.
21: **end for**

II. Method

A. Preliminary

Given an observation $p_t \in \mathbb{R}^{N_t \times 6}$ with N_t points extracted from RGB-D image at current timestep t, RISE [23] $\pi(p_t) = a_{t:t+T_a}$ learns a direct mapping from the current observation to future robot actions over a horizon of T_a (low-freq). Building upon RISE, our proposed force-aware policy, FoAR, incorporates high-frequency force/torque observations $f_{t-T_o:t} \in \mathbb{R}^{T_o \times 6}$ over a historical horizon of T_o (high-freq) as additional inputs.

B. Force-Aware Policy Design

Point Cloud Encoder. Following RISE [23], we employ sparse 3D encoder [4] with a shallow ResNet architecture [8] to process the point cloud $p_t \in \mathbb{R}^{N_t \times 6}$ into sparse point tokens $P_t \in \mathbb{R}^{N_p \times 512}$, where N_p represents the number of sparse point tokens after processing. A Transformer [22] with sparse point encodings [23] is then applied to these point tokens to generate a scene feature $h_t^s \in \mathbb{R}^{512}$.

Force/Torque Encoder. The force/torque observation $f_t \in \mathbb{R}^6$ is first processed through a 3-layer MLP to generate the corresponding force token $F_t \in \mathbb{R}^{512}$. These tokens over the past horizon $F_{t-T_o:t} \in \mathbb{R}^{T_o \times 512}$, being inherently time-series data in nature, are then encoded using a Transformer [22] with sinusoidal positional encodings applied along the temporal axis, resulting in a force feature $h_t^f \in \mathbb{R}^{512}$.

Feature Fusion. We introduce a future contact predictor $\phi(t) \in [0, 1]$ to guide the feature fusion process. Specifically, the fused feature h_t is calculated as follows:

$$h_t = \left[h_t^s; \phi(t) \cdot h_t^f + (1 - \phi(t)) \cdot h^*\right],$$

where h^* is a learnable embedding, and $[\cdot;\cdot]$ is the concatenation symbol.

Future Contact Predictor. As discussed in §II-A, we use current RGB image I_t and force/torque data $f_{t-T_o:t}$ as observation inputs to the predictor, since (1) using RGB images can make the predictor more lightweight given that it performs similarly with point clouds in contact state determination; (2) while force/torque data does not directly predict future contact, it helps correct the predictor when unexpected contact occurs.

Action Head. The fused feature h_t is then used as the conditioning input for the action denoising process [3, 10, 20] to generate robot end-effector actions by progressively refining noisy action trajectories.

Supervision. The generated action is supervised by ground-truth action in demonstration data via L2 loss $\mathcal{L}_{\text{action}}$ in the diffusion process. The ground-truth future contact state is automatically extracted from the demonstrations based on whether the force/torque data exceeds a threshold δ_{demo} within a surrounding time window around the current timestep, which supervises the future contact predictor through binary cross-entropy loss $\mathcal{L}_{\text{predictor}}$. The overall loss $\mathcal{L}$ is a linear combination of both terms:

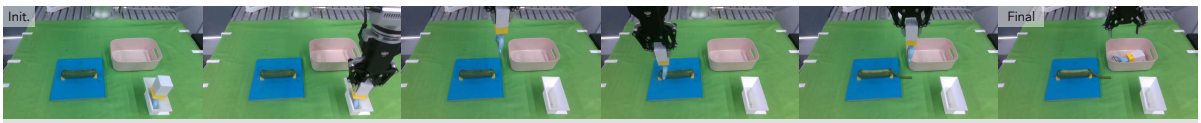
$$\mathcal{L} = \mathcal{L}_{\text{action}} + \alpha \mathcal{L}_{\text{predictor}},$$

where α is the weighting factor.

C. Reactive Control in Deployment

We introduce reactive control during deployment, as outlined in Alg. 1. Specifically, we threshold the predicted future contact probability ϕ from the contact predictor to determine whether the robot will make contact with the object and whether the predicted end-effector action needs to be adjusted using force/torque feedback. If ϕ exceeds the threshold δ_ϕ, indicating that the robot is in contact or will soon make contact with the object, the controller will check the current force/torque readings f_t, and correct the predicted robot actions if insufficient force/torque is detected. For action correction (Line 12-14 in Alg. 1), we estimate the future action direction based on the predicted action chunk and the current end-effector pose q_t, then adjust the predicted robot actions by a small step ϵ towards that direction.

Wiping*: use the eraser to wipe the whiteboard. The robot needs to (1) *grasp* the eraser, (2) *move* above a whiteboard with random figures, (3) *wipe* the figures on the whiteboard, and (4) *place* the eraser in the container after wiping. The ***Wiping (General) task allows arbitrary whiteboard orientations.**



***Peeling*: use the peeler to peel cucumber. The robot needs to (1) *grasp* the peeler, (2) *move* the peeler closer to the board, aligning above the cucumber, (3) *peel* the cucumber, and (4) *place* the peeler in the container after peeling.**

***Chopping*: use the knife to chop peppers. The robot needs to (1) *grasp* the knife, (2) *move* the knife closer to the board, aligning above the pepper, (3) steadily *cut* through the pepper, and (4) *place* the knife to the foam pad.**

Fig. 3: Tasks. We carefully design 3 challenging contact-rich tasks that focus on different aspects of the contact-rich manipulations. These tasks involve both non-contact phases and contact phases to evaluate the policy performance thoroughly.

Method	*Wiping*			*Wiping* (General)			*Peeling*		
	Score ↑	ASR (%) ↑		Score ↑	ASR (%) ↑		Score ↑	ASR (%) ↑	
		Grasp	Wipe		Grasp	Wipe		Grasp	Peel
ACT [28]	0.275	65	50	0.250	65	50	0.120	95	25
Diffusion Policy [3]	0.400	75	60	0.350	75	50	0.386	85	70
RISE [23]	0.500	**100**	75	0.500	90	80	0.377	**100**	50
RISE (force-token)	0.575	85	80	0.600	90	80	0.487	95	75
RISE (force-concat)	0.475	**100**	65	0.675	**100**	95	0.524	**100**	80
FoAR (3D-cls)	0.175	40	35	0.200	40	40	0.270	95	40
FoAR (*ours*)	0.875	**100**	**100**	0.850	**100**	**100**	0.756	**100**	**100**

TABLE I: Evaluation Results of the *Wiping* and *Peeling* Tasks. ASR denotes the action success rate, measuring the success rates of the robot in executing certain actions, regardless of the quality of the actions.

Design Choice			Score ↑	ASR (%) ↑	
F/T Freq.	*w.* Predictor	*w.* Reactive		Grasp	Wipe
100Hz		✓	0.650	**100**	85
100Hz	✓		0.650	**100**	80
2Hz	✓	✓	0.625	**100**	85
10Hz	✓	✓	0.800	**100**	**100**
100Hz	✓	✓	0.875	**100**	**100**

TABLE II: Ablation Results of the *Wiping* Task on Several Design Choices. We ablate our design choices on contact predictor, reactive control, and high-frequency force/torque sensing.

By incorporating reactive control during deployment, our FoAR policy can effectively handle uncertainties and dynamic changes in the environment, allowing the robot to adapt to real-world variations and achieve more reliable contact-rich manipulation performance.

III. Experiments

A. Setup

Platform. The system employs a Flexiv Rizon arm with Dahuan AG-95 gripper and OptoForce F/T sensor. Visual perception is provided by an Intel RealSense D435 camera. Computation is handled by an Intel i9-10900K/NVIDIA RTX 3090 workstation.

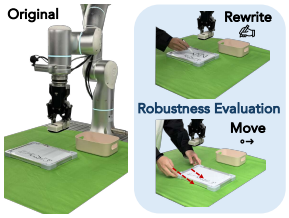
Tasks. As shown in Fig. 3, we design three challenging contact-rich tasks across two categories: surface force control (***Wiping*** and ***Peeling***) and instantaneous force impact (***Chopping***).

Baselines. We evaluate our proposed approach against five baseline methods, including the vision-based policy ACT [28], Diffusion Policy [3], RISE [23] and three ablation variants: *RISE (force-token)*, *RISE (force-concat)*, and *FoAR (3D-cls)*.

Metrics. Task performance is quantified through domain-specific measures: The ***Wiping*** task employs a three-tiered success metric (1.0 for complete cleaning, 0.5 for partial success, and 0 for failure) based on residual marker visibility. For the ***Peeling*** operation, effectiveness is measured by the ratio of removed vegetable skin relative to expert demonstration benchmarks. The ***Chopping*** evaluation combines segment count analysis with statistical consistency metrics, calculating both the mean and standard deviation of normalized segment lengths to assess cutting precision. All tasks additionally report fundamental action success rates (ASR) to verify task completion capability.

Protocols. For policy training, we collect 50 expert demonstrations for the ***Wiping*** and ***Peeling*** tasks, and 40 for the ***Chopping*** task. During evaluation, we run 20 trials per method for the ***Wiping*** and ***Peeling*** tasks, and 10 trials on the ***Chopping*** task.

Implementation. FoAR uses $T_o = 200$ to encode high-frequency (100Hz) force/torque data, corresponding to approximately 2 seconds of data. The dimensions of force tokens, scene feature h_t^s, force feature h_t^f, and learnable embedding h^* are all set to 512. For the future contact predictor, we utilize a ResNet18 [8] vision encoder and an MLP-based force encoder, followed by feature concatenation and a linear layer to output the probability ϕ. We combine the action loss and the predictor loss using $\alpha = 0.1$ during



Method	*Original*			*Rewrite*			*Move*			*Rewrite + Move*		
	Score ↑	ASR(%) ↑		Score ↑	ASR(%) ↑		Score ↑	ASR(%) ↑		Score ↑	ASR(%) ↑	
		Grasp	Wipe		Grasp	Wipe		Grasp	Wipe		Grasp	Wipe
RISE [23]	0.500	90	80	0.500	80	70	0.600	**100**	**100**	0.500	**100**	70
RISE (force-token)	0.600	90	80	0.450	90	90	0.500	90	80	0.600	**100**	**100**
FoAR (*ours*)	0.850	**100**	**100**	0.800	**100**	**100**	0.850	**100**	**100**	0.800	**100**	**100**

TABLE III: Robustness Evaluation Results of the *Wiping* (General) Task. The figure on the left illustrates the dynamic disturbances in the robustness evaluation. "Original" refers to vanilla evaluation with no disturbances.

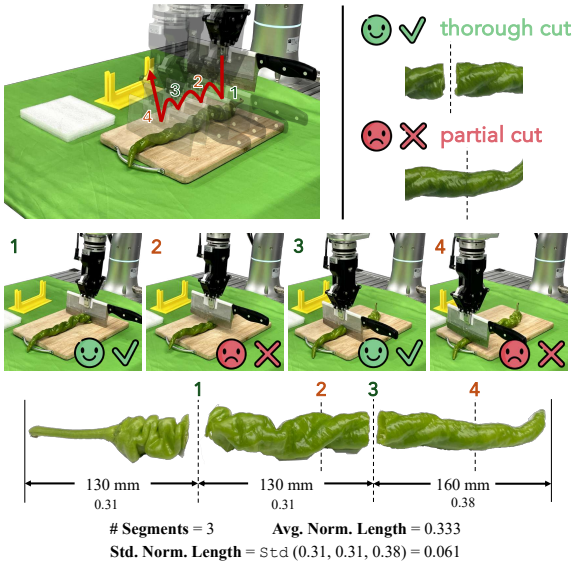


Fig. 4: Evaluation Metrics of the *Chopping* Task. We encourage the robot to divide the pepper into several uniform small segments, without segments sticking together due to partial cuts.

training. Other hyperparameters remain the same as RISE. For reactive control in deployment, we set the future contact probability threshold $\delta_\phi = 0.9$, force threshold $\delta_f = 8\text{N}$, torque threshold $\delta_t = 5\text{N}\cdot\text{m}$, and small step $\epsilon = 0.006\text{m}$.

*B. Surface Force Control Tasks: **Wiping** and **Peeling***

In surface force control tasks (***Wiping*** and ***Peeling***), the robot utilizes force/torque data to maintain consistent surface contact. As shown in Fig. 3, the ***Wiping*** task assesses the ability of the policy to maintain continuous and sustained contact, while the ***Peeling*** task emphasizes precision and sensitivity in manipulation. We report the evaluation results for the ***Wiping***, ***Wiping*** (General), and ***Peeling*** tasks in Table I. Our FoAR policy significantly outperform all baselines and variants.

*C. Instantaneous Force Impact Task: **Chopping***

The ***Chopping*** task evaluates the robot's ability to handle instantaneous force impacts, requiring precise force and torque control that vision data alone cannot provide [25]. The main challenge lies in accurately assessing the chopping as the knife's contact with the pepper and the chopping depth constantly change. The results in Tab. IV demonstrate that FoAR outperforms the baseline policy RISE, providing more reliable and controlled performance in the ***Chopping*** task.

Method	# Segments ↑	Norm. Length		ASR (%) ↑	
		Avg. ↓	Std. ↓	Grasp	Place
RISE [23]	1.8 ± 0.6	0.727	0.411	**100**	30
FoAR (*ours*)	**3.9** ± 0.9	**0.353**	**0.094**	**100**	**70**
Oracle (demonstration)	5.0 ± 0.0	0.200	0.056	100	100

TABLE IV: Evaluation Results of the *Chopping* Task. We also calculate the metrics of the demonstrations as an oracle for reference.

D. Ablations

Designing contact predictor, applying reactive control, and integrating high-frequency force/torque sensing enable the policy to perform more precise contact-rich manipulations. To illustrate the importance of these design, we take the ***Wiping*** task as an example. Ablation results can be found in Tab. II.

E. Robustness to Dynamic Disturbances

FoAR maintains consistent task performance under unexpected and dynamic environmental disturbances, demonstrating superior robustness and adaptability. As shown in Tab. III, FoAR successfully maintains consistent performance across all robustness evaluations, adapting to several dynamic environmental disturbances. While RISE also demonstrates strong generalization ability [23], its performance is limited by the absence of force/torque feedback. Built upon RISE, FoAR inherits this generalization ability while leveraging force/torque integration to achieve superior performance. In contrast, RISE (force-token) struggles in these complex scenarios, likely due to disturbances forcing the policy into non-contact phases, requiring action regeneration. Noise in force/torque data during these transitions further amplifies errors, hindering its effectiveness.

IV. Conclusion

In this paper, we propose FoAR, a force-aware reactive policy tailored for contact-rich robotic manipulation. By introducing a future contact predictor, the policy enables effective contact-guided feature fusion between force/torque and visual information, dynamically balancing the contribution of each modality based on future contact probability. This design not only enhances precision during contact phases but also maintains strong performance in non-contact phases. Additionally, the future contact probability further guides the reactive control strategy, improving policy performance even with simple position control.

References

[1] Anthony Brohan et al. "RT-1: Robotics Transformer for Real-World Control at Scale". In: *Robotics: Science and Systems*. 2023.

[2] Tao Chen et al. "Vegetable Peeling: A Case Study in Constrained Dexterous Manipulation". In: *arXiv preprint arXiv:2407.07884* (2024).

[3] Cheng Chi et al. "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion". In: *Robotics: Science and Systems*. 2023.

[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. "4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3075–3084.

[5] Open X-Embodiment Collaboration et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models". In: *IEEE International Conference on Robotics and Automation*. 2024, pp. 6892–6903.

[6] Siyuan Dong and Alberto Rodriguez. "Tactile-based Insertion for Dense Box-Packing". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019, pp. 7953–7960.

[7] Ruoxuan Feng et al. "Play to the Score: Stage-Guided Dynamic Multi-Sensory Fusion for Robotic Manipulation". In: *Conference on Robot Learning*. 2024.

[8] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[9] Minho Heo et al. "FurnitureBench: Reproducible Real-World Benchmark for Long-Horizon Complex Manipulation". In: *Robotics: Science and Systems*. 2023.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[11] Yifan Hou et al. "Adaptive Compliance Policy: Learning Approximate Compliance for Diffusion Guided Control". In: *arXiv preprint arXiv:2410.09309* (2024).

[12] Binghao Huang et al. "3D-ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing". In: *arXiv preprint arXiv:2410.24091* (2024).

[13] Moo Jin Kim et al. "OpenVLA: An Open-Source Vision-Language-Action Model". In: *arXiv preprint arXiv:2406.09246* (2024).

[14] Hao Li et al. "See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation". In: *Conference on Robot Learning*. 2022, pp. 1368–1378.

[15] Wenhai Liu et al. "ForceMimic: Force-Centric Imitation Learning with Force-Motion Capture System for Contact-Rich Manipulation". In: *arXiv preprint arXiv:2410.07554* (2024).

[16] Zeyi Liu et al. "ManiWAV: Learning Robot Manipulation from In-the-Wild Audio-Visual Data". In: *Conference on Robot Learning*. 2024.

[17] Jared Mejia et al. "Hearing Touch: Audio-Visual Pretraining for Contact-Rich Manipulation". In: *arXiv preprint arXiv:2405.08576* (2024).

[18] Octo Model Team et al. "Octo: An Open-Source Generalist Robot Policy". In: *Robotics: Science and Systems*. 2024.

[19] Branden Romero et al. "Eyesight Hand: Design of a Fully-Actuated Dexterous Robot Hand with Integrated Vision-based Tactile Sensors and Compliant Actuation". In: *arXiv preprint arXiv:2408.06265* (2024).

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: *The International Conference on Learning Representations*. 2021.

[21] Markku Suomalainen, Yiannis Karayiannidis, and Ville Kyrki. "A Survey of Robot Manipulation in Contact". In: *Robotics and Autonomous Systems* 156 (2022), p. 104224.

[22] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[23] Chenxi Wang et al. "RISE: 3D Perception Makes Real-World Robot Imitation Simple and Effective". In: *arXiv preprint arXiv:2404.12281* (2024).

[24] Yansong Wu et al. "TacDiffusion: Force-domain Diffusion Policy for Precise Tactile Manipulation". In: *arXiv preprint arXiv:2409.11047* (2024).

[25] Shangning Xia et al. "CAGE: Causal Attention Enables Data-Efficient Generalizable Robotic Manipulation". In: *arXiv preprint arXiv:2410.14974* (2024).

[26] Kelin Yu et al. "MimicTouch: Leveraging Multi-Modal Human Tactile Demonstrations for Contact-Rich Manipulation". In: *Conference on Robot Learning*. 2024.

[27] Andy Zeng et al. "Transporter Networks: Rearranging the Visual World for Robotic Manipulation". In: *Conference on Robot Learning*. 2020, pp. 726–747.

[28] Tony Z. Zhao et al. "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware". In: *Robotics: Science and Systems*. 2023.

[29] Brianna Zitkovich et al. "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control". In: *Conference on Robot Learning*. 2023, pp. 2165–2183.