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 [41]. Such tasks, including assembly [17, 48], wiping [21, 32], and peeling [8, 31], are inherently challenging due to the complex dynamics of force and precise control required. Unlike simple pick-and-place operations [50], 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 advancing robotic dexterity, enabling more versatile, autonomous, and interactive robot systems.

In recent years, significant progress has been made in vision-based robotic manipulation policies [5, 10, 12, 25, 37, 44, 47, 52, 54]. 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. This limitation hinders the robot's ability to perceive contact states and understand physical interactions, thus constraining its manipulation capabilities, as illustrated in Fig. 1 (left).

To address the limitations of pure vision-based policies, recent approaches have incorporated additional modalities such as audio [15, 30, 32, 35], tactile [15, 22, 30, 39], and force/torque [9, 28, 46, 53] into the policy framework. These multi-modal policies offer promising avenues for advancing robotic manipulation by providing richer feedback about interactions, enabling robots to handle contact-rich tasks with greater precision and adaptability.

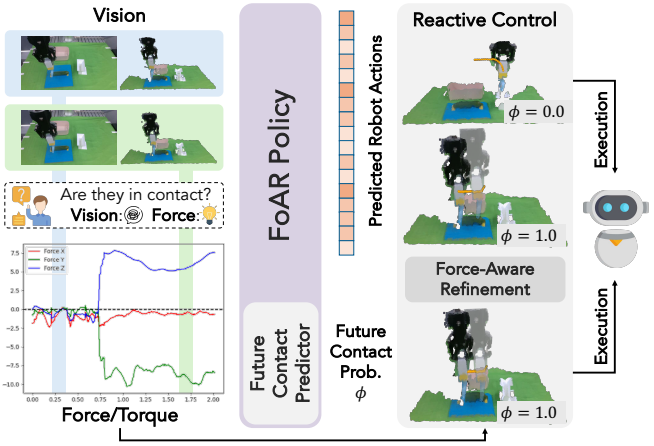


Fig. 1: Overview of the FoAR Policy for Contact-Rich Robotic Manipulations. Vision alone struggles to distinguish contact from non-contact states in contact-rich tasks, underscoring the need for integrating force/torque information. Our FoAR policy combines vision and force/torque inputs to predict robot 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.

Nevertheless, audio and similar indirect sensing modalities are typically vulnerable to background noise [32], complicating signal processing and reducing reliability in real-world applications. Additionally, they often fail to deliver detailed information about contact dynamics between robots and objects, constraining their effectiveness in tasks that require precise manipulation. Although tactile sensing provides direct contact information, it faces unique challenges due to the wide variety of available sensor types. For example, camera-based tactile sensors [27, 49] are highly heterogeneous, making it difficult to standardize the tactile perception results [51], while magnetic-based tactile sensors [4, 42] often encounter inconsistency issues during replacements [3], adding further complexity to their use.

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, leveraging this information allows policies to sense and adapt to contact interactions in real time, thereby

* 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

enhancing the precision and control of manipulation tasks. While prior studies [21, 31, 46] 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.

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 [44], 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. Additionally, FoAR demonstrates exceptional robustness, maintaining stable performance in three evaluation scenarios with unexpected dynamic disturbances, highlighting its adaptability and resilience in real-world applications.

II. Related Works

A. Integrating Force/Torque Perception in Manipulation

Force/torque perception is critical for enabling robots to interact effectively with the environment, particularly in manipulations that demand precise control and accurate feedback. By measuring the forces and torques applied to the robot, sensors offer valuable data on contact states, helping the robot perform delicate, contact-rich manipulations [7].

Early research leveraged force/torque feedback for low-level control strategies [2, 20, 33], enabling precise control in contact-rich tasks but often overlooking its potential for high-level decision-making. More recently, advancements have broadened the application of force/torque perception in robot learning. For example, methods such as [1, 21, 24] enhance vision-based policies [10, 52] by incorporating force/torque inputs and predefined stiffness outputs for compliance control. Others have extended the diffusion policy [10] into the force domain to predict contact wrenches for hybrid force/position control [31], feedforward forces for impedance control [46], and desired forces for admittance control [53].

However, these approaches often only emphasize contact phases by assuming the object is already grasped [1, 24, 46, 53] or fixed to the robot [21, 31], bypassing the impact of noisy force/torque readings during non-contact phases. Other works [6, 26] employ torque data for bilateral control but rely on leader-follower teleoperation frameworks [52], limiting their adaptability to different data collection setups.

B. Contact-Rich Robotic Manipulation

Contact-rich manipulation has been extensively studied due to its relevance in both manufacturing and daily life. It involves enabling robots to perform complex tasks that require precise control during physical interactions with the environment [41]. In the past, researchers developed classical force control methods [19, 34, 38, 45] for assembly tasks, laying the foundation for contact-rich manipulation control techniques. However, these approaches are often limited by their reliance on precise models and predefined strategies.

Recent advances in learning-based methods have greatly expanded robots' capabilities in contact-rich manipulation. Reinforcement learning-based approaches [23, 28, 29, 36, 48] enable robots to learn complex tasks through interaction, but they often struggle with sim-to-real transfer due to discrepancies in visual observations and force/torque feedback, limiting their performance in real-world tasks. Several imitation learning studies seek to improve the abilities of the robot in contact-rich manipulation abilities by incorporating auxiliary modalities like audio [15, 30, 32, 35], tactile [13, 15, 22, 30, 39], and force/torque [6, 21, 24, 31, 46, 53]. A key challenge in multimodal policies lies in effectively processing and integrating different modalities within the policy framework, ensuring that the information from each modality is applied to the relevant task phases.

By leveraging the force/torque modality, we propose learning a future contact probability to guide the multimodal feature fusion. This approach allows the force/torque information to enhance the contact phases of the task while preventing noisy data from interfering with other phases, leading to improved performance in various contact-rich manipulation tasks compared to previous fusion methods.

III. Method

A. Preliminary

Given an observation $p_t \in \mathbb{R}^{N_t \times 6}$ at current timestep t, RISE [44] $\pi(p_t) = a_{t:t+T_a}$ learns a direct mapping from the current observation to future robot actions over a horizon of T_a. 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 as additional inputs. Notice that T_o represents the history horizon for high-frequency force/torque data, typically sampled at about 100Hz, while T_a denotes the action horizon for future predictions, which operates at a lower frequency like 10Hz.

While vision-based policies [10, 37, 44, 47] have demonstrated success in simple contact-rich tasks, we argue that force/torque information is vital for more complex scenarios. Taking the determination of contact states as an example, Fig. 1 (left) shows that visual differences in either RGB images or point clouds before and after contact are minimal, making it difficult to determine contact states. In contrast, force/torque data provides clear and reliable indicators of contact, highlighting its critical role in such tasks. As a result, incorporating high-frequency force/torque data complements point cloud observations, enabling more accurate and robust decision-making in contact-rich manipulations.

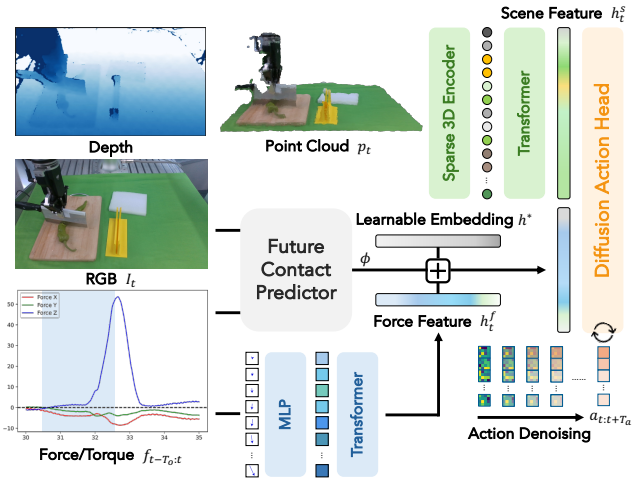


Fig. 2: FoAR Architecture. FoAR consists of a point cloud encoder [44], a force/torque encoder, a future contact predictor, and a diffusion action head [10]. The scene features and force features are fused under the guidance of the future contact predictor.

Algorithm 1 FoAR Inference with Reactive Control

buffer.clear();
contact_buffer.clear(); ▷ clear the temporal ensemble buffer.
for timestep $t \leftarrow 0$ to $N_{\text{max}} - 1$ **do**
  if t mod $N_{\text{inference}} = 0$ **then** ▷ inference time step.
    $p_t, I_t, f_{t-T_o:t}, q_t \leftarrow$ ***agent***.perception;
    $\phi, a_{t:t+T_a} \leftarrow \text{FoAR}(p_t, f_{t-T_o:t}, I_t)$;
    if $\phi < \delta_\phi$ **then** ▷ non-contact phase.
      buffer.add($a_{t:t+T_a}$);
    else ▷ contact phase.
      if force(f_t) $< \delta_f$ and torque(f_t) $< \delta_t$ **then**
        ▷ insufficient force/torque detected.
        $\mathbf{d} \leftarrow \text{avg}(a_{t:t+T_f}).\text{pos} - q_t.\text{pos}$;
        $a_{t:t+T_a}.\text{pos} \leftarrow a_{t:t+T_a}.\text{pos} + \epsilon \cdot \mathbf{d}/\|\mathbf{d}\|_2$;
        ▷ update actions towards predicted direction.
      end if
      contact_buffer.add($a_{t:t+T_a}$);
    end if
  end if
  $a_t \leftarrow$ ***buffer***.get(t) **if** $\phi < \delta_\phi$ **else** ***contact_buffer***.get(t);
  agent.execute(a_t); ▷ execute the end-effector action.
end for

B. Force-Aware Policy Design

Point Cloud Encoder. Following RISE [44], we employ sparse 3D encoder [11] with a shallow ResNet architecture [16] to process the point cloud $p_t \in \mathbb{R}^{N_t \times 6}$ into sparse point tokens $P_t \in \mathbb{R}^{N_p \times 512}$. A Transformer [43] with sparse point encodings [44] 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 [43] with sinusoidal positional encodings applied along the temporal axis, resulting in a force feature $h_t^f \in \mathbb{R}^{512}$.

Feature Fusion. Previous studies on multimodal feature fusion in robotic manipulation have explored approaches such as direct concatenation of features [28, 31] and processing multimodal tokens through Transformers [9, 30, 32, 37]. However, such simple fusion methods often lead to the noisy force modality interfering with the non-contact phases of the task. Instead, 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. In other words, the future contact predictor dynamically adjusts the weight of the force feature h_t^f in the fusion process, ensuring that the force data is strongly emphasized during contact phases while minimizing its impact during non-contact phases by blending it with a neutral embedding h^*. This approach allows the policy to more effectively utilize multimodal information without introducing interference from irrelevant data.

Future Contact Predictor. The future contact predictor takes the current observations as inputs and outputs the probability that contact will occur in the *future* steps. This probability is used to modulate the fusion of the visual and force modalities, allowing the model to emphasize force data when contact is likely to occur and reduce its influence during non-contact phases. As discussed in §III-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 [10, 18, 40] 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:

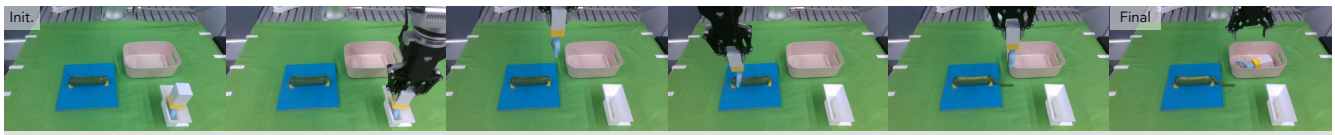
$$\mathcal{L} = \mathcal{L}_{\text{action}} + \alpha \mathcal{L}_{\text{predictor}},$$

where α is the weighting factor.

C. Reactive Control in Deployment

Prior literature has explored various control strategies for contact-rich manipulation, such as admittance control [53], compliance control [21, 24, 34], and hybrid force/position control [31, 38]. These approaches often require additional parameters, such as stiffness and contact force 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.

In contrast, we demonstrate that our proposed future contact predictor enables the robot to perform accurate, force-feedback-driven manipulation in contact-rich tasks even using simple end-effector position control, eliminating the need for complex parameter tuning and prediction.

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. Different temporal ensemble buffers [52] are used for contact and non-contact phases to avoid mutual interference while ensuring smooth trajectory execution.

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.

IV. Experiments

During the experiments, we intend to answer the following research questions:

- ***Q1***: Does integrating force/torque information improve policy performance and manipulation accuracy in contact-rich tasks, particularly in real-world scenarios where such tasks involve multiple phases with varying demands on precision and contact interactions?
- ***Q2***: Is the feature fusion module of the FoAR policy more effective than other variants in terms of integrating force/torque information?
- ***Q3***: Does reactive control during deployment enhance the policy's ability to perform contact-rich actions?
- ***Q4***: Can FoAR maintain consistent task performance under unexpected environmental disturbances?

A. Setup

Platform. The experimental platform consists of a Flexiv Rizon robotic arm with a Dahuan AG-95 gripper, and an OptoForce force/torque sensor mounted between the flange and the gripper. The robot operates within a 45cm $\times$ 60cm $\times$ 40cm workspace. An Intel RealSense D435 RGBD camera located in front of the robot workspace is used for scene perception. All devices are linked to a workstation with an Intel Core i9-10900K CPU and an NVIDIA RTX 3090 GPU for both data collection and evaluation.

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***). These tasks require different capabilities in terms of the direction, intensity, and precision of applied contact forces. Moreover, these tasks are designed to have both non-contact phases and contact phases for thorough evaluations. For the ***Wiping*** task, we design two variants: one with a fixed orientation of the whiteboard, and another that allows arbitrary orientations, denoted as ***Wiping*** (General).

Baselines. We evaluate our proposed approach against five baseline methods, including the vision-based policy RISE [44] and three ablation variants: (1) *RISE (force-token)*: incorporates encoded force/torque information as additional tokens within the RISE transformer, akin to [9, 30, 32,

Method	*Wiping*			*Wiping* (General)			*Peeling*		
	Score ↑	ASR (%) ↑		Score ↑	ASR (%) ↑		Score ↑	ASR (%) ↑	
		Grasp	Wipe		Grasp	Wipe		Grasp	Peel
RISE [44]	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	-	-	-	0.524	**100**	80
FoAR (3D-cls)	0.175	40	35	-	-	-	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. For the ***Wiping*** (General) task, we only select the best-performing ablation variant on the ***Wiping*** task, RISE (force-token), for evaluation.

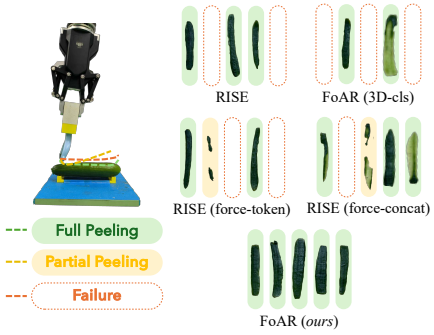


Fig. 4: Qualitative Results of the *Peeling* Task.

37]; (2) *RISE (force-concat)*: directly concatenates the force feature with the vision feature for action generation; (3) *FoAR (3D-cls)*: uses scene features h_t^s directly in the future contact predictor, instead of a separate image encoder.

Metrics. For all tasks, we calculate the action success rate (referred to as ASR) to assess the policy's ability to meet basic action requirements, regardless of action quality. For the ***Wiping*** task, the score is assigned to 1 for a fully wiped whiteboard, 0.5 for partial wiping, and 0 for no erasure. For the ***Peeling*** task, the score is calculated based on the proportion of peeled cucumber skin to the total cucumber length, normalized by the average proportion in the demonstration data (0.778). For the ***Chopping*** task, we aim to let the robot use the knife to divide the pepper into several uniform small segments. Therefore, we focus on the number of segments, as well as the mean and standard deviation of the normalized lengths (defined as the proportion of each segment's length to the total length of the pepper), providing a comprehensive assessment of chopping precision and consistency, as shown in Fig. 5.

Protocols. For policy training, we collect 50 expert demonstrations for the ***Wiping*** and ***Peeling*** tasks, and 40 for the ***Chopping*** task through haptic teleoperation [14]. During evaluation, we run 20 trials per method for the ***Wiping*** and ***Peeling*** tasks, and 10 trials each only for FoAR and RISE [44] on the ***Chopping*** task to conserve resources. Objects are randomly placed in the workspace, while ensuring similar positions across methods for fair comparisons.

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 [16] 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 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. A key challenge arises from the variability in tool grasp positions (*e.g.*, top, bottom, center, or off-center), requiring adaptive control to adjust for changes in the grasp. 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.

FoAR significantly outperforms baselines in surface force control tasks by integrating force/torque information to enhance manipulation accuracy and contact consistency in surface force control tasks (*Q1*). We report the evaluation results for the ***Wiping***, ***Wiping*** (General), and ***Peeling*** tasks in Table I. Our proposed method, FoAR, achieves the highest scores of 0.875, 0.850, and 0.756 for the ***Wiping***, ***Wiping*** (General), and ***Peeling*** tasks, respectively, significantly outperforming all baseline and variant methods. FoAR attains 100% success rates in both grasping the tool and performing the contact-rich operations (wiping and peeling) in all tasks, demonstrating its ability to maintain continuous and precise contact regardless of grasp position of the tool (eraser and peeler). In contrast, the pure vision-based policy RISE struggles with these contact-rich operations due to the lack of force/torque feedback, leading to inaccurate position control, which reflects the difficulty in maintaining consistent contact stemming from absence of force/torque perceptions. The qualitative results of the ***Peeling*** task in Fig. 4 further support these findings, showcasing that FoAR achieves more consistent and effective performance compared to the baselines, which often result in partial peelings or failures.

The feature fusion module of FoAR enhances capabilities in contact-rich operations while maintaining strong performance during non-contact phases, surpassing several variants in integrating force/torque information (*Q2*). We report the evaluation results for these variants in Table I and Fig. 4. RISE (force-token) and RISE (force-concat) exhibit similar or slightly better performance compared to the pure vision-based policy RISE, suggesting that incorporating force/torque data as an additional input does provide some benefits. However, the key factor lies in how these inputs

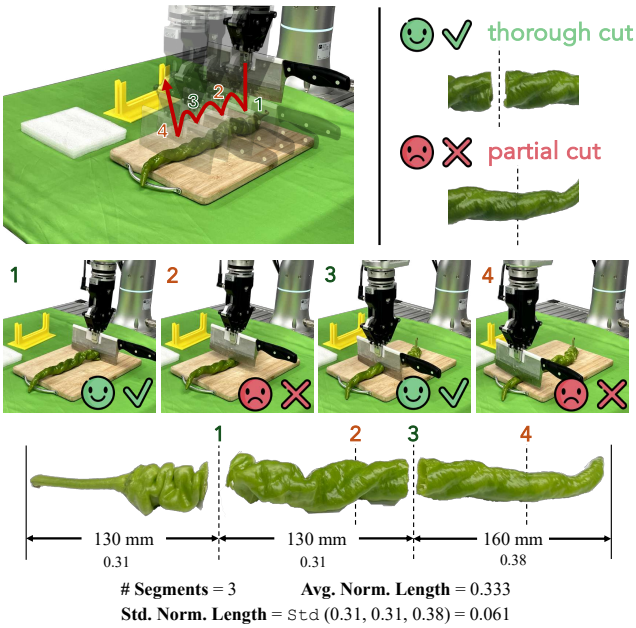


Fig. 5: Evaluation Metrics of the *Chopping* Task. We encourage the robot to divide the pepper into several uniform small segments, without segments sticking together due to partial cuts.

Method	# Segments ↑	Norm. Length		ASR (%) ↑	
		Avg. ↓	Std. ↓	Grasp	Place
RISE [44]	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 II: Evaluation Results of the *Chopping* Task. We also calculate the metrics of the demonstrations as an oracle for reference.

are effectively leveraged in the policy. Simply integrating force/torque tokens into the policy transformer or concatenating force features with vision features not only fails to fully leverage force/torque information but also risks introducing noisy force/torque data during non-contact phases, which can interfere with the policy's decision-making process and thus negatively impacting overall performance, *e.g.*, leading to lower grasp action success rates in both tasks for RISE (force-token). On the contrary, FoAR demonstrates strong performance in both contact and non-contact phases, highlighting the effectiveness of our feature fusion module over these variants in utilizing force/torque data.

Separating the future contact predictor from the policy backbone is crucial to avoid disruption (*Q2*). As shown in Table I, the FoAR (3D-cls) variant even significantly underperforms the RISE baseline. This variant employs a shared sparse 3D encoder for both the future contact predictor and the policy backbone. We suspect that the visual features required by each component differ substantially. For example, in the ***Wiping*** task, the future contact predictor focuses on whether the eraser is positioned above the whiteboard, whereas the policy requires detailed information like the precise locations of objects and the end-effector position. Consequently, sharing a single vision encoder may cause conflicting attention and potential interference, disrupting both components and reducing their effectiveness.

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 [47]. The main challenge lies in accurately assessing the chopping as the knife's contact with the pepper and the chopping depth constantly change.

Vision alone is insufficient for ensuring smooth chops, highlighting the necessity of force/torque feedback for precise control and improved policy performance (*Q1*). The results in Tab. II demonstrate that FoAR outperforms the baseline policy RISE, providing more reliable and controlled performance in the ***Chopping*** task. It achieves over double the number of segments (3.9 *v.s.* 1.8) and a lower averaged normalized segment length (0.353 *v.s.* 0.727) with reduced segment variability (standard deviation of 0.094 *v.s.* 0.411), indicating better performance in chopping the pepper into smaller, more uniform segments.

D. Ablations on Reactive Control

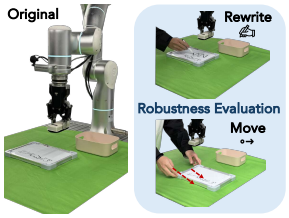
Reactive control in deployment enables the policy to perform more precise contact-rich actions (*Q3*). To illustrate the importance of reactive control during policy deployment, we use the ***Wiping*** (General) task as an example. The results in Table III demonstrate that reactive control is essential for achieving optimal policy performance. Notably, our proposed reactive control relies on the predicted future contact probability from the policy, highlighting the effective co-design of the FoAR policy and the reactive control mechanism. Consequently, in all experiments, reactive control is applied to FoAR-based methods to ensure high performance.

Method	Score ↑	ASR(%) ↑	
		Grasp	Wipe
FoAR *wo.* Reactive Control	0.650	**100**	80
FoAR *w.* Reactive Control	0.850	**100**	**100**

TABLE III: Ablation Results of the *Wiping* (General) Task on Reactive Control.

E. Robustness to Dynamic Disturbances

To further assess the adaptability of our model FoAR under more challenging and varied conditions, we develop three robustness evaluations for the ***Wiping*** (General) task: (1) *Rewrite*: write new random figures on the wiped area after robot wiping; (2) *Move*: move the whiteboard to a different position after robot wiping; (3) *Rewrite + Move*: combine the previous two dynamic disturbances, *i.e.*, move the whiteboard to a different position, and write new random figures on the wiped area after robot wiping. These evaluations introduce disturbances during task execution to assess how well the methods can adjust to new conditions.



Method	*Original*			*Rewrite*			*Move*			*Rewrite + Move*		
	Score ↑	**ASR**(%) ↑		**Score** ↑	**ASR**(%) ↑		**Score** ↑	**ASR**(%) ↑		**Score** ↑	**ASR**(%) ↑	
		Grasp	Wipe		Grasp	Wipe		Grasp	Wipe		Grasp	Wipe
RISE [44]	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 IV: 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.

FoAR maintains consistent task performance under unexpected and dynamic environmental disturbances, demonstrating superior robustness and adaptability (*Q4*). As shown in Tab. IV, FoAR achieves scores of 0.800, 0.850, and 0.800 for the ***Rewrite***, ***Move***, and ***Rewrite + Move*** robustness evaluations, respectively, while maintaining 100% action success rates. It can adapt to newly written random figures in erased areas, adjust its strategy to changes in whiteboard position, and handle both challenges simultaneously. As a strong vision-based baseline, RISE also exhibits decent generalization ability and handles these variations without performance degradation [44], but its overall performance is constrained by the absence of force/torque feedback. Built upon RISE, FoAR successfully inherits the strong generalization ability, consistently detecting and responding to dynamic environmental changes while maintaining high performances. Its effective integration of force/torque information elevates performance to a higher level compared to RISE. Conversely, RISE (force-token) struggles to handle such complex scenarios and experiences a performance drop. We hypothesize that the unexpected disturbances force the policy to return to non-contact phases, requiring action sequence re-generations. The inherent noise in force/torque data during these phases further exacerbates errors in the re-generated sequences, hindering its effectiveness.

V. 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. Extensive experiments demonstrate the superior performance of FoAR in contact-rich tasks that require sustained and precise contact, such as wiping, peeling, and chopping. In the future, we plan to integrate advanced control strategies, such as compliance control and hybrid force/position control, into the FoAR policy to further enhance its performance. We also aim to extend this approach to dual-arm robots or humanoid robots for more complex contact-rich manipulation tasks.

Acknowledgement

We would like to thank Chenxi Wang for helpful discussions, Yiming Wang and Shangning Xia for their help during the data collection process.

References

[1] Malek Aburub et al. "Learning Diffusion Policies from Demonstrations For Compliant Contact-rich Manipulation". In: *arXiv preprint arXiv:2410.19235* (2024).

[2] Cristian Camilo Beltran-Hernandez et al. "Learning Force Control for Contact-Rich Manipulation Tasks With Rigid Position-Controlled Robots". In: *IEEE Robotics and Automation Letters* 5.4 (2020), pp. 5709–5716.

[3] Raunaq Bhirangi et al. "Anyskin: Plug-and-play skin sensing for robotic touch". In: *arXiv preprint arXiv:2409.08276* (2024).

[4] Raunaq M. Bhirangi et al. "ReSkin: Versatile, Replaceable, Lasting Tactile Skins". In: *Conference on Robot Learning*. 2021, pp. 587–597.

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

[6] Thanpimon Buamanee et al. "Bi-ACT: Bilateral Control-Based Imitation Learning via Action Chunking with Transformer". In: *arXiv preprint arXiv:2401.17698* (2024).

[7] Max Yiye Cao, Stephen Laws, and Ferdinando Rodriguez y Baena. "Six-Axis Force/Torque Sensors for Robotics Applications: A Review". In: *IEEE Sensors Journal* 21.24 (2021), pp. 27238–27251.

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

[9] Yizhou Chen et al. "Visuo-Tactile Transformers for Manipulation". In: 2022.

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

[11] 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.

[12] 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.

[13] 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.

[14] Hao-Shu Fang et al. "RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot". In: *IEEE International Conference on Robotics and Automation*. 2024, pp. 653–660.

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

[16] 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.

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

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

[19] Neville Hogan. “Impedance Control: An Approach to Manipulation”. In: *Journal of Dynamic Systems, Measurement, and Control* 107 (1985), pp. 1–24.

[20] Yifan Hou and Matthew T Mason. “Robust Execution of Contact-Rich Motion Plans by Hybrid Force-Velocity Control”. In: *IEEE International Conference on Robotics and Automation*. 2019, pp. 1933–1939.

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

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

[23] Mrinal Kalakrishnan et al. “Learning Force Control Policies for Compliant Manipulation”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2011, pp. 4639–4644.

[24] Tatsuya Kamijo, Cristian C Beltran-Hernandez, and Masashi Hamaya. “Learning Variable Compliance Control from a Few Demonstrations for Bimanual Robot with Haptic Feedback Teleoperation System”. In: *arXiv preprint arXiv:2406.14990* (2024).

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

[26] Masato Kobayashi, Thanpimon Buamanee, and Takumi Kobayashi. “ALPHA-α and Bi-ACT Are All You Need: Importance of Position and Force Information/Control for Imitation Learning of Unimanual and Bimanual Robotic Manipulation with Low-Cost System”. In: *arXiv preprint arXiv:2411.09942* (2024).

[27] Mike Lambeta et al. “DIGIT: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor With Application to In-Hand Manipulation”. In: *IEEE Robotics and Automation Letters* 5.3 (2020), pp. 3838–3845.

[28] Michelle A. Lee et al. “Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks”. In: *IEEE International Conference on Robotics and Automation*. IEEE, 2019, pp. 8943–8950.

[29] Sergey Levine, Nolan Wagener, and Pieter Abbeel. “Learning contact-rich manipulation skills with guided policy search”. In: *IEEE International Conference on Robotics and Automation*. 2015, pp. 156–163.

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

[31] 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).

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

[33] Emanuele Magrini, Fabrizio Flacco, and Alessandro De Luca. “Control of Generalized Contact Motion and Force in Physical Human-Robot Interaction”. In: *IEEE International Conference on Robotics and Automation*. 2015, pp. 2298–2304.

[34] Matthew T Mason. “Compliance and Force Control for Computer Controlled Manipulators”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 11.6 (1981), pp. 418–432.

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

[36] Michael Noseworthy et al. “FORGE: Force-Guided Exploration for Robust Contact-Rich Manipulation under Uncertainty”. In: *arXiv preprint arXiv:2408.04587* (2024).

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

[38] MH Raibert and JJ Craig. “Hybrid Position/Force Control of Manipulators”. In: *Journal of Dynamic Systems, Measurement, and Control* 103.2 (1981), pp. 126–133.

[39] 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).

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

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

[42] Tito Pradhono Tomo et al. “A New Silicone Structure for uSkin - A Soft, Distributed, Digital 3-Axis Skin Sensor and Its Integration on the Humanoid Robot iCub”. In: *IEEE Robotics and Automation Letters* 3.3 (2018), pp. 2584–2591.

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

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

[45] Daniel E. Whitney. “Historical Perspective and State of the Art in Robot Force Control”. In: *IEEE International Conference on Robotics and Automation*. 1985, pp. 262–268.

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

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

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

[49] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. “GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force”. In: *Sensors* 17.12 (2017).

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

[51] Jialiang Zhao et al. “Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks”. In: *Conference on Robot Learning*. 2024.

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

[53] Bo Zhou et al. “Admittance Visuomotor Policy Learning for General-Purpose Contact-Rich Manipulations”. In: *arXiv preprint arXiv:2409.14440* (2024).

[54] 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.

APPENDIX

A. Implementation Details

Data Processing. Following RISE [44], we create the point cloud from a single-view RGB-D image captured by a global camera. Then both the input point clouds and the output actions are aligned in the same camera coordinate. The point cloud is cropped based on the pre-defined robot workspace (notice that the tabletop points remain after cropping). The coordinates are normalized to $[-1, 1]$ based on the robot workspace. The gripper width is also normalized to $[-1, 1]$ according to the gripper width range.

Point Cloud Encoder. We implement sparse 3D encoder using MinkowskiEngine [11] with a voxel size of 5mm. The sparse 3D encoder outputs a set of 512-dimensional point feature vectors. The transformer [43] contains 4 encoding blocks and 1 decoding block, with $d_{\text{model}} = 512$ and $d_{\text{ff}} = 2048$. The readout token has a dimension of 512.

Force/Torque Encoder. The high-frequency force/torque observation of the last $T_o = 200$ steps (approximately 2 seconds given the frequency of 100Hz) is encoded via a 3-layer MLP of dimension $(64, 128, 512)$. We use the same transformer architecture as the point cloud encoder to process these force/torque tokens with sinusoidal positioning encoding along the temporal axis. The readout token has a dimension of 512.

Future Contact Predictor. We utilize a ResNet18 [16] vision encoder and a two-layer MLP of dimension $(128, 512)$ to process image and force/torque inputs respectively. The outputs are concatenated and passed through a linear layer to compute the future contact probability ϕ. The ground-truth future contact state t is generated from the force/torque data within the time window $[t - 2\text{s}, t + 2\text{s}]$ and checks whether force/torque value exceeds the predefined force and torque thresholds. The thresholds may differ across tasks and can be easily determined manually from the collected demonstrations. The ground-truth future contact states are then used to supervise the future contact predictor.

Action Head. A CNN-based diffusion head [10] is employed with 100 denoising iterations for training and 20 iterations for inference using DDIM [40] scheduler. FoAR predicts the future action of $T_a = 20$ steps.

Training. FoAR is trained on 2 NVIIDA A100 GPUs with a batch size of 240, an initial learning rate of 3×10^{-4}, and a warmup step of 2000. The learning rate is decayed by a cosine scheduler. During training, we apply the same point cloud augmentations as RISE [44], and we also leverage color jittering for improved robustness. The weighting factor α between the future contact predictor loss and the action loss is set as 0.1.

Deployment. We apply the reactive control during deployment, combining with simple end-effector position control (*i.e.*, Line 20 in Alg. 1 is sent to the end-effector position controller of the robot). No advanced control strategies like compliance control, admittance control, and hybrid force/torque control are used in this paper. The future contact probability threshold δ_ϕ is set to 0.9, the force threshold δ_f is set to 8N, the torque threshold δ_t is set to $5\text{N} \cdot \text{m}$. The motion direction is calculated based on the predicted future $T_f = 5$ action steps, and we set the small step $\epsilon = 0.006\text{m}$.

B. Additional Ablation

We conduct an additional ablation experiment by replacing the transformer in the force/torque encoder with a simple MLP, as in the future contact predictor.

Method	Score ↑	ASR(%) ↑	
		Grasp	Peel
RISE [44]	0.293	**100**	50
FoAR (MLP)	0.426	**100**	75
FoAR	0.588	**100**	**100**

TABLE V: Ablation Results of the *Peeling* Task on Different Force/Torque Encoders.

The results shown in Tab. V illustrate that the simple MLP encoder cannot effectively capture the temporal features of the force/torque information, resulting in inferior performance compared to the transformer-based encoder. However, it still outperforms RISE by incorporating force/torque information for contact-rich manipulations.