Predictive Force Conditioning for Contact-Rich Imitation Learning

Yigit Korkmaz[1], Keegan Go[2], Gautam Salhotra[3], Erdem Bıyık[1], Ajinkya Jain[2], Stefan Schaal[2]
[1]University of Southern California [2]Intrinsic [3]Google DeepMind

***Abstract*—Many fine-grained manipulation tasks require force feedback for reliable execution. Despite its importance, most robot learning frameworks rely primarily on visual and proprioceptive observations, often overlooking force–torque (FT) signals. In contrast, FT sensing is readily available on many robotic platforms and provides dense feedback at substantially higher frequencies than other sensing modalities. In this work, we investigate how force–torque observations can be integrated into robot learning frameworks while preserving the rich information they provide. We propose a two-stage policy architecture that first predicts future force–torque measurements and subsequently conditions action generation on these predictions. By explicitly modeling future interaction dynamics, our approach grounds action generation in anticipated physical feedback rather than relying solely on immediate observations. Through preliminary experiments, we analyze the attention distributions of different policy architectures across visual and force modalities, providing insight into how force information is utilized during action generation. Our results show that the proposed framework encourages utilization of force information and achieves higher success rates than existing vision-only and force-aware baselines.**

I. Introduction

Humans naturally rely on multiple sensory modalities when performing fine-grained manipulation tasks. During contact-rich interactions, such as inserting a connector or aligning tightly fitting components, visual feedback alone is often insufficient to determine task progress. Instead, humans continuously monitor contact forces and their subtle variations, dynamically shifting their attention toward force feedback when it becomes more informative than vision. Once contact is established, force signals can reveal information about alignment, friction, and interaction dynamics that may not be directly observable from images alone.

Despite the importance of force feedback, most robot learning approaches rely primarily on visual observations and proprioception. Recent methods like Diffusion Policy (DP) [1] and Action Chunking Transformers (ACT) [2] have demonstrated impressive performance across a wide range of manipulation tasks, yet they often struggle in contact-rich settings where successful execution depends on interpreting physical interactions with the environment. Meanwhile, force-torque (FT) sensing is readily available on many modern robot arms, such as Franka Panda and KUKA LBR iiwa, providing dense measurements of contact dynamics at higher frequencies than visual observations. This raises a natural question: how can robot learning systems effectively leverage force information without sacrificing the rich contextual information provided by vision?

Incorporating FT signals into policy learning presents several challenges. First, force measurements are typically collected at substantially higher frequencies than images and proprioceptive data. Naively subsampling these signals discards much of the information captured by the sensors, potentially removing critical cues about contact events and interaction dynamics. As a result, effective integration of FT sensing requires architectures that can preserve the high-frequency structure of force observations while maintaining their relationship with other modalities.

Second, force signals are inherently sparse in their informativeness. For large portions of an episode, the robot may not be in contact with the environment, causing force readings to remain near zero and contribute little information beyond sensor noise. However, when contact occurs, these signals become highly informative and often provide stronger guidance than visual observations. This imbalance creates a challenging optimization problem: policies trained on multimodal observations may learn to rely primarily on visual features and effectively ignore force information, since visual cues are consistently available throughout training. Consequently, successful integration of FT sensing requires mechanisms that not only preserve force information but also encourage policies to attend to it when it matters most.

To address these challenges, we propose **P**redictive-**F**orce-**C**onditioning Policies (PFC), a two-stage framework for incorporating force-torque sensing into robot policy learning. Given visual observations, and force-torque readings, PFC first predicts future force measurements and subsequently conditions action generation on these predictions along with the input observations. By explicitly modeling future interaction forces, our method grounds action generation in anticipated contact dynamics rather than relying solely on instantaneous observations. Furthermore, PFC leverages the full high-frequency FT signal via a frequency-aware encoder, avoiding information loss caused by subsampling. The framework also incorporates explicit cross-attention between visual and force embeddings to capture dependencies between the two modalities.

We evaluate PFC on a preliminary contact-rich peg insertion task and compare it against several force-aware and vision-only baselines. In addition, we analyze attention distributions across visual and force modalities, providing insight into how different architectures leverage force feedback during

policy execution. Our results show that conditioning action generation on predicted future forces leads to higher task success rates and better utilization of force information than existing approaches.

II. Related Work

Recent advances in robot imitation learning have led to increasingly capable visuomotor policies, ranging from behavior cloning and autoregressive transformer architectures to diffusion-based policies and, more recently, vision-language-action (VLA) models [1, 2, 3, 4]. These approaches have demonstrated strong performance across a wide range of manipulation tasks using visual observations and proprioceptive state information [5]. However, many real-world manipulation tasks require reasoning about physical interactions with the environment, where force feedback plays a critical role in improving robustness, adaptability, and execution stability [6, 7].

To address this limitation, several recent works have explored incorporating FT sensing into robot learning frameworks. Most existing approaches treat force measurements as an additional observation modality and focus on improving the fusion of visual and force information [8, 9, 10, 11, 12]. Liu et al. [8] encourage policies to attend more strongly to force signals through curriculum-based visual perturbations and Zhang et al. [10] use information bottleneck objectives. He et al. [13] explicitly predict contact and non-contact phases to adaptively regulate the fusion of visual and force representations. Xue et al. [12] employs a two-stage framework in which force information is used to refine high-level actions generated by a diffusion policy. Other approaches explore mixture-of-experts architectures [11], adaptive compliance control strategies with learned dynamic gains [14], and specialized architectures for handling the heterogeneous frequencies of visual and force observations, including cross-attention mechanisms and causal convolutional networks for force encoding [9, 15]. While these methods improve multimodal fusion, they primarily utilize force sensing as an additional observation stream rather than as an explicit representation of future interaction dynamics.

A separate line of work investigates force-aware action generation. Chen et al. [16] leverage measured contact forces from kinesthetic demonstrations to reconstruct force-informed action targets that better capture interaction dynamics. Liu et al. [17] jointly predict force and motion parameters and executes them through a hybrid force–position controller for contact-rich manipulation. In contrast to these approaches, we propose a two-stage policy architecture in which future FT measurements are explicitly predicted and subsequently used to condition action generation. Rather than treating force solely as an observation or control target, our framework uses predicted future force interactions as an intermediate representation that grounds policy actions in anticipated contact dynamics.

III. Preliminaries

We consider an imitation learning setting in which a robot learns a policy from a dataset of expert demonstrations. Each demonstration consists of a sequence of observations and actions, $\tau = \{(o_1, a_1), (o_2, a_2), \ldots, (o_T, a_T)\}$, where o_t denotes the observation at time step t and a_t denotes the corresponding expert action.

The observation o_t is composed of different modalities v_t and F_t, where v_t represents visual observations captured by one or more cameras and F_t corresponds to FT measurements obtained from a wrist-mounted force–torque sensor.

Unlike visual observations, FT measurements are typically collected at substantially higher frequencies. Therefore, during a policy execution window, multiple FT observations may be available for each visual observation.

$$F_t = f_t^{(1)}, f_t^{(2)}, \ldots, f_t^{(K)}$$

denote the sequence of FT measurements collected between two consecutive visual observations, where K depends on the relative sampling frequencies of the sensors.

IV. Method

A key challenge in incorporating FT sensing into robot learning is effectively using high-frequency force signals while maintaining consistency with visual observations. FT measurements are often weakly informative during free-space motion but become critical during contact-rich interactions, where subtle changes in force can reveal information that is difficult to infer from vision alone. Our goal is to design a policy architecture that can leverage this information without discarding the rich temporal structure present in FT observations.

To this end, we introduce PFC, a transformer-based two-stage policy learning framework that explicitly models future interaction dynamics through force prediction. Given visual observations and high-frequency FT measurements, PFC first predicts future FT values and subsequently conditions action generation on these predictions. By reasoning about expected contact forces, the policy can generate actions that are grounded in future interaction dynamics rather than relying only on instantaneous observations. Furthermore, PFC preserves the full temporal resolution of FT measurements through a dedicated causal convolutional encoder and employs explicit cross-attention between force and visual tokens to capture dependencies across modalities.

A. Observation Encoding and Cross-Attention

At each policy timestep, the robot receives visual observations and a sequence of high-frequency FT measurements. While visual observations are sampled at the policy frequency, FT measurements are collected at a substantially higher rate. Consequently, each policy timestep is associated with a sequence of K force observations, $F_t = f_t^{(1)}, f_t^{(2)}, \ldots, f_t^{(K)}$, where each $f_t^{(i)} \in \mathbb{R}^6$ contains three force and three torque measurements recorded by the wrist-mounted FT sensor.

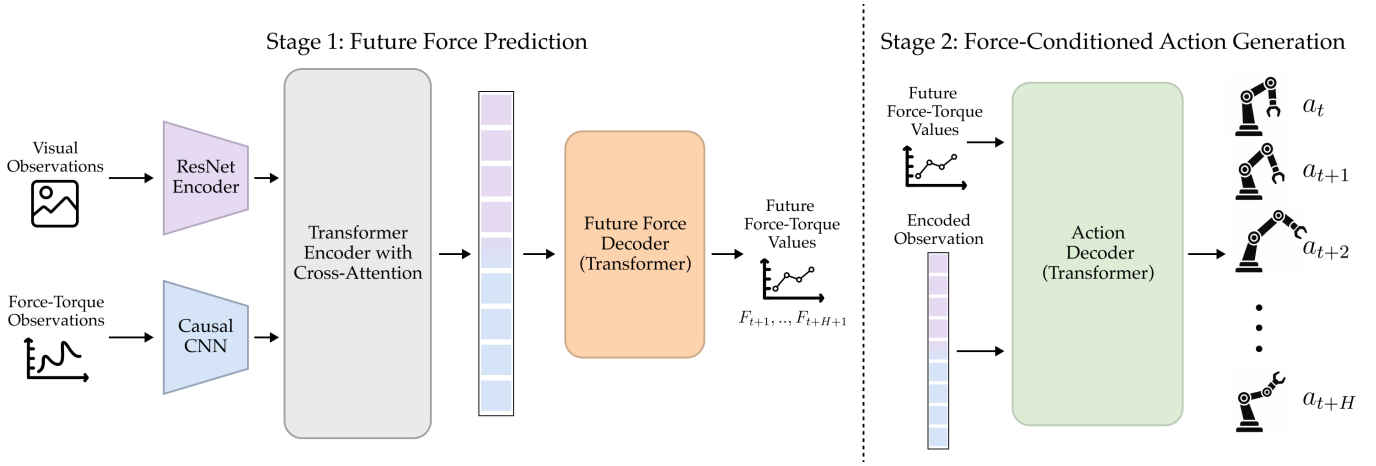

Fig. 1: **Overview of Predictive Force Conditioning (PFC).** Visual observations and high-frequency force–torque (FT) measurements are encoded using modality-specific encoders and fused through cross-attention. The model first predicts a future FT trajectory, which is then used as additional conditioning for action generation. This two-stage structure encourages the policy to reason about anticipated contact dynamics during contact-rich manipulation.

Visual observations from each camera are encoded using a frozen ResNet [18] backbone to extract visual feature representations. To prevent the high-dimensional visual embeddings from dominating the multimodal fusion process, the extracted features are projected into a shared latent space of dimension d through a lightweight MLP. We additionally introduce learned camera-specific positional embeddings to distinguish observations originating from different viewpoints.

To encode the high-frequency FT stream, we employ a causal convolutional network [19]. Unlike simple subsampling strategies, the causal CNN preserves the temporal structure of FT measurements while extracting local force dynamics relevant for contact-rich manipulation. The resulting FT embeddings are projected into the same latent space as visual features and augmented with learned frequency-aware positional embeddings that encode the temporal ordering of FT samples within each policy timestep.

While visual and force observations provide complementary information, their relevance varies throughout task execution. To capture interactions between these modalities, we introduce explicit bidirectional cross-attention between visual and FT representations. Let $Z_v \in \mathbb{R}^{N_v \times d}$ and $Z_f \in \mathbb{R}^{N_f \times d}$ denote the visual and force embeddings, respectively, where N_v and N_f correspond to the number of visual and force tokens.

Force-conditioned visual representations are computed as

$$\mathrm{softmax}\left(\frac{Q_v K_f^\top}{\sqrt{d}}\right) V_f,$$

where

$$Q_v = Z_v W_Q^v, \qquad K_f = Z_f W_K^f, \qquad V_f = Z_f W_V^f.$$

Similarly, visual-conditioned force representations are obtained as

$$\mathrm{softmax}\left(\frac{Q_f K_v^\top}{\sqrt{d}}\right) V_v,$$

where

$$Q_f = Z_f W_Q^f, \qquad K_v = Z_v W_K^v, \qquad V_v = Z_v W_V^v.$$

This bidirectional attention mechanism enables the model to explicitly associate visual observations with corresponding force patterns and vice versa. The resulting cross-modal representations are concatenated with the original modality embeddings and passed to subsequent transformer layers.

B. Future Force Prediction

Motivated by the observation that future force measurements provide a compact representation of upcoming interaction dynamics, the first stage of PFC predicts future FT values from the encoded visual and force observations. Specifically, given the multimodal representation at time t, the model predicts a horizon of future force measurements, $\{\hat{f}_{t+1}, \hat{f}_{t+2}, \ldots, \hat{f}_{t+H+1}\}$, where H denotes the prediction horizon. Note that here, each $\hat{f}_{t+i} \in \mathbb{R}^6$

The force prediction objective encourages the model to learn task-relevant interaction dynamics, such as impending contact events, alignment corrections, and insertion forces. Rather than treating FT measurements solely as observations, this stage learns a representation of how physical interactions are expected to evolve in the near future.

C. Force-Conditioned Action Generation

The second stage conditions policy generation on both the current multimodal observations and the predicted future force trajectory. Specifically, the predicted force embeddings are concatenated with the multimodal latent representation and provided as additional conditioning to the policy transformer.

Formally, the policy is defined as

$$\pi(a_{t:t+H} \mid v_t, F_t, \{\hat{f}_{t+1}, \hat{f}_{t+2}, \ldots, \hat{f}_{t+H+1}\}),$$

where the predicted future forces serve as an intermediate representation of anticipated interaction dynamics. By conditioning actions on future force predictions, PFC encourages

the policy to reason about the physical consequences of its actions before they occur, leading to more robust behavior in contact-rich manipulation tasks.

D. Training Objectives

To demonstrate the generality of our framework, we instantiate PFC using two commonly adopted policy architectures: an ACT-style autoregressive policy and a diffusion-based policy. In both variants, the observation encoder and transformer backbone are shared between force prediction and action generation.

In the ACT-based variant, the model predicts a horizon of future force measurements in parallel from the shared transformer representation. The predicted force embeddings are then provided as additional conditioning to the action prediction head. The force prediction module is trained using an ℓ_2 regression objective, $L_{\text{force}} = \frac{1}{H}\sum_{h=1}^{H}\left|\hat{f}_{t+h} - f_{t+h}\right|_2^2$, while the action prediction head is optimized using the standard ℓ_1 reconstruction objective, $L_{\text{action}} = |\hat{a}_t - a_t|_1$.

For the diffusion-based variant, both future force prediction and action generation are formulated as denoising processes. Following standard diffusion training, Gaussian noise is independently added to the future force trajectory and action sequence. The model then predicts the corresponding noise terms, ϵ_F and ϵ_A, for force prediction and action generation, respectively. The predicted future force trajectory is subsequently used as conditioning for the action denoising network. The diffusion objective is defined as

$$\mathcal{L}_{\text{force}} = \mathbb{E}\left[|\epsilon_F - \hat{\epsilon}_F|_2^2\right],$$
$$\mathcal{L}_{\text{action}} = \mathbb{E}\left[|\epsilon_A - \hat{\epsilon}_A|_2^2\right],$$

The overall training objective is given by

$$\mathcal{L} = \mathcal{L}_{\text{action}} + \lambda\mathcal{L}_{\text{force}},$$

where λ controls the relative importance of the force prediction objective.

A challenge when training the action generation module is that force prediction quality is initially poor. Conditioning action generation on inaccurate force predictions early in training can therefore destabilize learning. To address this issue, we employ a staged conditioning strategy. During the first 80% of training iterations, the action generation module is conditioned on the ground-truth future force trajectory. Once the force prediction module has learned a reasonable estimate of future interaction dynamics, we switch to using the predicted future force values for the remaining 20% of training. This strategy allows the policy to benefit from accurate force conditioning early in training while ensuring that, by the end of optimization, the policy operates under the same conditions encountered at inference time. The complete PFC framework is illustrated in Figure 1.

V. Experiments

We perform a series of experiments on a contact-rich round peg insertion task using a NIST board [20] to evaluate three aspects of our approach. First, we compare PFC against vision-only policies and prior force-aware baselines on task performance. Second, we analyze whether the proposed two-stage force prediction framework encourages policies to utilize force information more effectively than existing methods. Finally, we evaluate the quality of the learned future force predictions and analyze their role in action generation.

A. Experimental Setup

All experiments are performed on a 7-DOF KUKA iiwa manipulator [21] equipped with a wrist-mounted FT sensor. We focus specifically on the final phase of the manipulation process, peg insertion after coarse alignment has already been achieved. Accordingly, each episode begins with the peg positioned near the target hole and in contact with the board surface, requiring the policy to reason about contact dynamics in order to successfully complete the insertion. The setup includes four RGB cameras: two external corner-view cameras and two wrist-mounted cameras, each operating at 10 Hz. The robot is controlled at 100 Hz, while the FT sensor records measurements at 1 kHz. Consequently, approximately 100 force readings are collected between two consecutive image observations. For all baseline methods, we provide only the most recent FT measurement at each policy timestep, which is the common practice in prior work. In contrast, PFC utilizes the full sequence of 100 FT measurements, which are encoded using a causal CNN to preserve their temporal structure. The action space consists of target joint positions for the robot arm.

We collect a dataset of 50 demonstrations using a GELLO setup [22]. Furthermore, we provide the operator with a real-time monitoring interface that visualizes the measured force and torque values at the end effector. This allows the demonstrator to incorporate force information during teleoperation while collecting demonstrations.

We compare PFC against the following baselines:

- **ACT** [2]: An action-chunking transformer policy that predicts a sequence of future actions from visual observations.
- **Diffusion Policy (DP)** [1]: A diffusion-based visuomotor policy that generates actions through iterative denoising.
- **FACTR** [8]: A force-aware policy learning framework that encourages stronger utilization of force signals through visual perturbations during training.
- **Joint Force–Action Estimation**: An ablation that jointly predicts future force values and actions using a shared objective, without the proposed two-stage force-conditioning mechanism.

We evaluate two instantiations of PFC. The first, PFC-A, employs an autoregressive action and force prediction objective, while the second, PFC-D, utilizes a diffusion-based generation objective. Both variants share the same multimodal encoder, force prediction module, and force-conditioned action generation framework.

Two-stage force conditioning improves task performance. Table I reports the success rates of all methods on the peg insertion task over 20 evaluation trials. Both variants

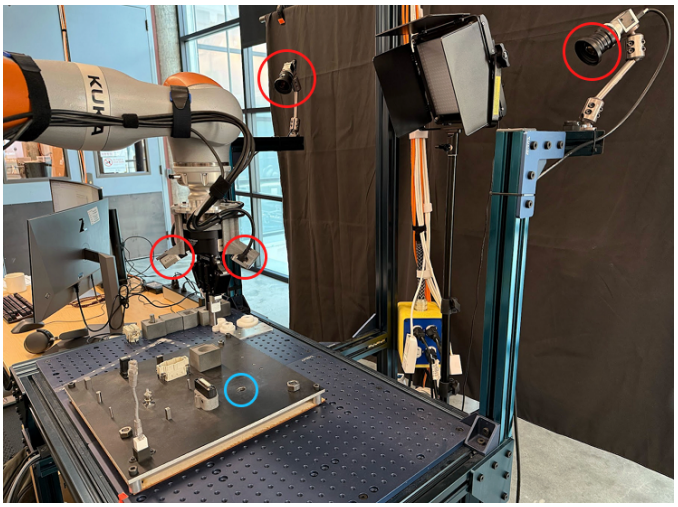

Fig. 2: **Experiment setup.** There are four cameras in the setup. Two external, and two wrist cameras, each operating at 10 Hz. The built-in FT sensor in the end-effector streams values at 1 kHz. The task is to insert round peg into the correct hole.

TABLE I: Success rates on the peg insertion task. Results are reported over 20 evaluation trials.

Method	Successes / 20
ACT	5/20
ACT + FT	3/20
DP	8/20
DP + FT	6/20
FACTR	4/20
Joint Force–Action Estimation	6/20
PFC-A	10/20
PFC-D	**12/20**

of PFC achieve the highest success rates, with PFC-A getting 10/20 and PFC-D achieving 12/20 successes. In contrast, both ACT and DP obtain limited benefit from the addition of force observations, and in some cases performance decreases relative to their vision-only counterparts. This behavior suggests that naively incorporating force measurements as an additional input may not be sufficient and can even introduce overfitting in low-data regimes. FACTR improves performance by explicitly encouraging force utilization through curriculum-based visual perturbations, yet both PFC variants outperform it.

PFC encourages better utilization of force. To better understand how different policies utilize various observation modalities, we analyze the attention distributions of the first decoder layer during policy execution. Following prior work [8], we visualize the average cross-attention from action tokens to visual and force tokens throughout the rollout. As shown in Figure 3, action generation in PFC places substantially more attention on force tokens than on visual tokens, which includes both current and future force tokens. In contrast, the force prediction module attends to both visual and force representations, suggesting that future force estimation relies on information from both modalities. For comparison, we also visualize the attention distribution of an ACT policy that receives force measurements as an additional observation modality, and a variant that jointly predicts forces and actions.

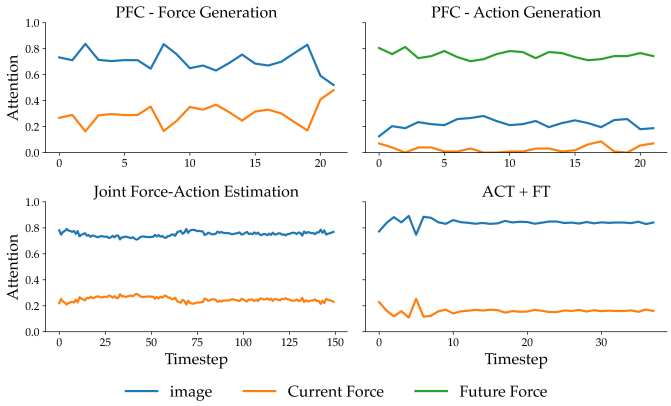

Fig. 3: Average cross attention of the action tokens to the force or vision tokens of the first decoder layer during policy rollout.

We observe that ACT mostly attends to visual tokens, indicating that simply providing force observations does not lead to effective utilization of force information. Similarly, the joint force–action prediction variant continues to rely primarily on visual observations.

Future force predictions capture contact dynamics. Finally, we evaluate the quality of the future force predictions generated by PFC. Figure 4 compares the predicted and measured force values during a successful evaluation rollout. We observe that the predicted force values closely follow the overall trends and key transitions present in the measured forces, indicating that the model successfully captures task-relevant interaction dynamics. Combined with the attention analysis, these results suggest that the predicted force trajectory provides a useful intermediate representation for action generation.

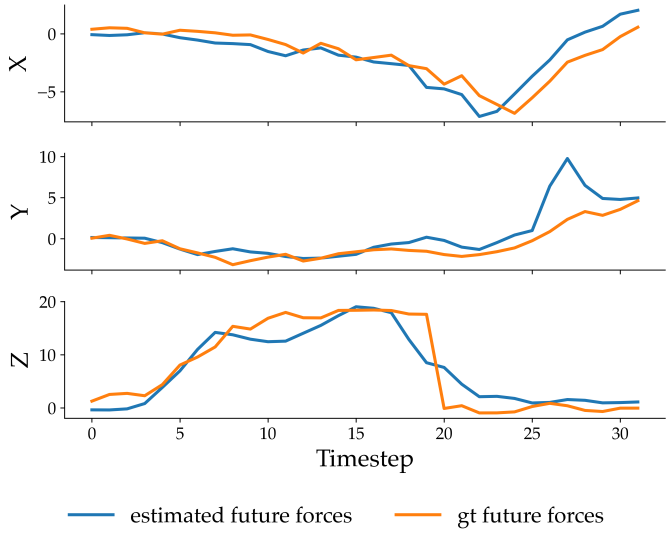

Fig. 4: Comparison of ground-truth vs. predicted future force values during a successful evaluation rollout.

Taken together, these results indicate that future force prediction is a promising mechanism for incorporating force feedback into robot learning framework. Our preliminary experiments demonstrate that conditioning actions on predicted

force values leads to improved force utilization and better performance on contact-rich manipulation, motivating further investigation of this approach on a broader range of tasks and policy architectures.

VI. Limitations and Future Work

While our preliminary results are promising, several limitations remain. First, we observe that the future force prediction module can overfit to the training demonstrations. Since action generation is conditioned on the predicted forces, out-of-distribution force predictions directly hurt policy performance. Future work should investigate more robust force prediction strategies, potentially using play data or other unlabeled interaction data that can improve force prediction without requiring expert actions. Second, our evaluation is limited to a single contact-rich insertion task. Future work should do broader evaluations and ablation studies to better understand the individual contributions of high-frequency force encoding, explicit vision-force cross-attention, and future force conditioning across a wider range of manipulation tasks and baselines. Finally, our experiments utilize the built-in force-torque sensor of the KUKA iiwa, and it remains unclear how well the proposed framework generalizes to standalone FT sensors or other robotic platforms.

References

[1] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[4] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Robotics: Science and Systems (RSS)*, 2024.

[5] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," in *Conference on Robot Learning (CoRL)*, 2025.

[6] M. T. Mason, "Compliance and force control for computer controlled manipulators," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, no. 6, pp. 418–432, 2007.

[7] N. Hogan, "Impedance control: An approach to manipulation," in *1984 American control conference*. IEEE, 1984, pp. 304–313.

[8] J. J. Liu, Y. Li, K. Shaw, T. Tao, R. Salakhutdinov, and D. Pathak, "Factr: Force-attending curriculum training for contact-rich policy learning," in *Robotics: Science and Systems*, 2025.

[9] G. Lee, Y. Lee, K. Kim, S. Lee, S. Noh, S. Back, and K. Lee, "Manipforce: Force-guided policy learning with frequency-aware representation for contact-rich manipulation," *arXiv preprint arXiv:2509.19047*, 2025.

[10] Y. Zhang, Y. Wang, X. Sun, K. Huang, Z. Xu, J. Ji, Z. Che, J. Tang, and J. Sun, "Craft: Adapting vla models to contact-rich manipulation via force-aware curriculum fine-tuning," *arXiv preprint arXiv:2602.12532*, 2026.

[11] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, W. Zhang, and C. Lu, "ForceVLA: Enhancing VLA models with a force-aware moe for contact-rich manipulation," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

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

[13] Z. He, H. Fang, J. Chen, H.-S. Fang, and C. Lu, "Foar: Force-aware reactive policy for contact-rich robotic manipulation," *IEEE Robotics and Automation Letters*, 2025.

[14] 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," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[15] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Conference on Robot Learning*, 2022.

[16] 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.

[17] 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," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[20] K. Kimble, K. Van Wyk, J. Falco, E. Messina, Y. Sun, M. Shibata, W. Uemura, and Y. Yokokohji, "Benchmarking protocols for evaluating small parts robotic assembly systems," *IEEE robotics and automation letters*, 2020.

[21] KUKA AG, "Lbr iiwa: The sensitive lightweight robot," Webpage, 2026. [Online]. Available: https://www.kuka.com

[22] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," 2023.