FARM: Force-Aware Robotic Manipulation with Tactile-Conditioned Diffusion Policies

Erik Helmut[1], Niklas Funk[1], Tim Schneider[1], Cristiana de Farias[1], Jan Peters[1,2,3,4]

Abstract— **Contact-rich manipulation requires precise force control, yet many imitation-learning approaches treat visuotactile feedback as a passive observation rather than an explicit control target. In this work, we present Force-Aware Robotic Manipulation (FARM), an imitation learning framework that leverages high-dimensional tactile data to define a force-based action space. Using a modified version of the handheld Universal Manipulation Interface (UMI) gripper equipped with a GelSight Mini tactile sensor, we collect human demonstrations and deploy them on a matching actuated gripper. During policy rollouts, the proposed FARM diffusion policy jointly predicts robot pose, grip width, and grip force. FARM outperforms several baselines across high-force, low-force, and dynamic force adaptation tasks, demonstrating the advantages of force-grounded, high-dimensional tactile observations and a force-based control space. The codebase and design files are open-sourced and available at `https://tactile-farm.github.io`.**

I. INTRODUCTION

Humans naturally regulate grasp forces through touch, applying just enough pressure to prevent an object from slipping [1], [2]. In robotics, the selection of an appropriate grasping force has long been recognized as a crucial issue [3], especially when handling fragile or deformable objects, such as fruits or eggs. While tactile sensing provides vital information regarding slip and forces [4], effectively leveraging these signals for direct force control remains challenging.

Imitation learning has emerged as an efficient strategy for learning robotic manipulation from human demonstrations [5]. However, current tactile-integrated imitation learning approaches typically treat tactile sensing as a passive observation modality rather than an active component of the action space [6]. Consequently, contact forces often remain an uncontrolled byproduct of kinematic gripper commands.

In this work, we address this gap by introducing Force-Aware Robotic Manipulation (FARM), an imitation learning framework that integrates tactile feedback into the action space, while also leveraging a high-dimensional force-profile as an observation modality.

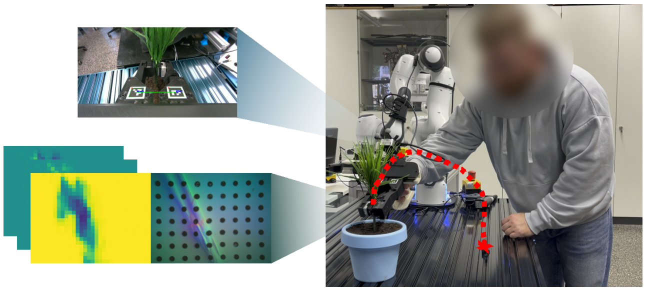

Fig. 1: Data-collection setup. Right: Expert demonstration using the adapted UMI gripper. Top-left: In-hand RGB view with ArUco markers for grip width measurement. Bottom-left: GelSight Mini tactile image and corresponding FEATS force estimates visualizing contact interactions.

METHOD

A. Gripper Hardware and Data Collection

To bridge the gap between human demonstrations and robot execution, we utilize two geometrically matched grippers: an adapted hand-held UMI gripper [7] for data collection and a custom-built *Actuated UMI* gripper for deployment. Both feature an Intel RealSense D405 for in-hand vision and a GelSight Mini tactile sensor at the fingertip.

During demonstrations, we record in-hand RGB images, gripper pose (via OptiTrack), and high-resolution tactile images. We employ FEATS [8] to extract estimates of shear and normal force distributions from the sensors' raw tactile images (see Fig. 1).

B. FARM Diffusion Policy

We extend the diffusion policy introduced by Chi et al. [9], to include a tactile-informed action space. As shown in Fig. 2 the FARM diffusion policy is conditioned on a multi-modal observation $\mathbf{O}_t$ consisting of:

1) **An RGB image**, captured by a Intel RealSense D405 camera mounted on the gripper.
2) **The grip width**, calculated as the Euclidean distance between the centers of the ArUco markers on the two fingers.
3) **Tactile feedback**, represented by a 3-channel image encoding the full force distributions extracted from each GelSight Mini tactile image using a pretrained and fixed FEATS model. Additionally, the total normal force is included as a scalar value, computed by integrating over the discretized normal force distribution.
4) **The gripper pose**, consisting of 3D position coordinates and a 6D rotation feature representation [10].

Corresponding author: Erik Helmut. Email: erik@robot-learning.de.

[1]Department of Computer Science, Technical University of Darmstadt [2]German Research Center for AI (DFKI) [3]Centre for Cognitive Science, Technical University of Darmstadt [4]Hessian Center for Artificial Intelligence (hessian.AI), Darmstadt

This work has been partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG) and the French Research Agency, l'Agence Nationale de Recherche (ANR), through the project *Aristotle* (ANR-21-FAI1-0009-01) and the EU's Horizon Europe project ARISE (Grant no.: 101135959).

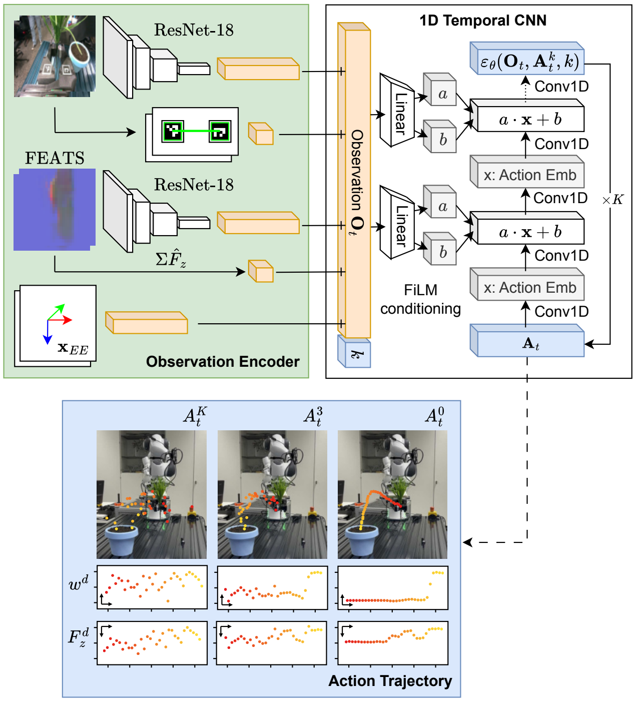

Fig. 2: FARM diffusion policy architecture. Visual, proprioceptive, and tactile observations are encoded and provided as input to a 1D temporal CNN with FiLM conditioning. The model predicts action trajectories for end-effector pose, grip width, and grip force, enabling closed-loop force control of the gripper during manipulation.

The model jointly predicts a trajectory of end-effector poses $\mathbf{x}_{EE}^d$, target grip widths w_d, and target grip forces F_z^d.

C. Policy Deployment and Gripper Control

To deploy learned policies on the *Actuated UMI* gripper, we implement a dual-model control strategy that switches between grip width control and force control based on the current interaction phase. During phases when there is no contact, the gripper width will be set with position control. If both the target force F_z^d and estimated force $\hat{F}_z$ are below -0.5 N, the system assumes that the robot is in contact with the object and switches to closed-loop force control.

D. Baselines

To evaluate the contribution of tactile feedback and force control in our FARM framework, we compare it against three baseline strategies: a *Force-Aware* baseline that utilizes only scalar normal force without FEATS force distributions, a *Tactile-Aware* baseline that processes the raw tactile images directly but lacks explicit force actions, and a *Vision-Only* baseline that relies exclusively on the in-hand RGB image and proprioception with binary gripper commands.

RESULTS

We evaluate the FARM framework against force-aware, tactile-aware, and vision-only baselines across three real-world tasks: plant insertion, grape picking, and screw tightening (see Fig. 3 and Fig. 4).

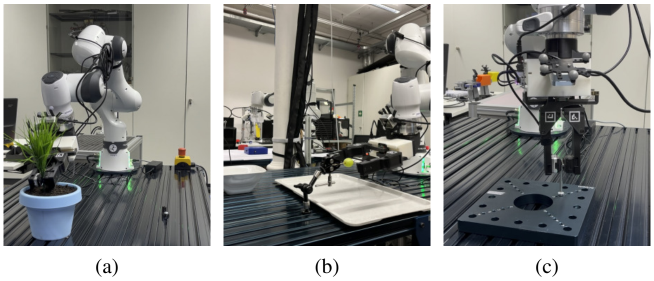

Fig. 3: Experimental tasks. (a) Plant insertion: planting into a soil-filled pot; (b) Grape picking: delicate grasping and detachment from a toothpick; (c) Screw tightening: using an Allen key.

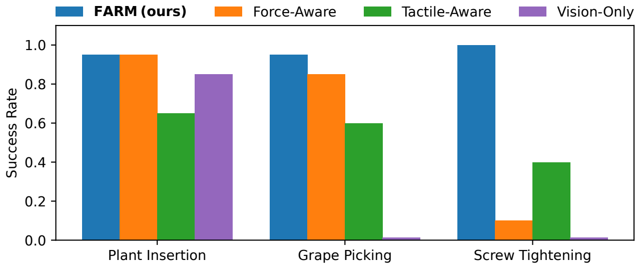

Fig. 4: Task success rates. Comparison of FARM against all baselines across three tasks, evaluated over 20 rollouts per task.

In the plant insertion and grape picking tasks, FARM (95% success) and the force-aware baseline ($85-95\%$) perform similarly, suggesting that scalar force signals are sufficient for tasks with static force demands. The vision-only baseline crushes the grapes, while the tactile-aware baseline (60%) frequently drops objects due to lack of direct force control.

The screw tightening task highlights the advantages of FARM's high-dimensional force-distribution representation. It achieved 100% success, while the force-aware baseline (10%) lacks information about the shear force distribution needed to maintain tool alignment. Furthermore, the tactile-aware baseline (40%) fails to reliably interpret the complex contact state from raw images, and the vision-only baseline (0%) lacks the tactile feedback to sense screw resistance. Quantitatively, FARM demonstrates the highest fidelity to human demonstrations in the force domain, yielding a Wasserstein-1 distance of 0.75 N, which is significantly lower than all baselines (ranging 1.66 N – 5.05 N).

CONCLUSION

This work introduces FARM, a visuotactile-conditioned diffusion policy that treats tactile feedback as both an observation and an explicit action space. By predicting target grip forces alongside robot poses, FARM enables delicate, contact-rich manipulation. Our results demonstrate that high-dimensional tactile observations, combined with a force-based action space, significantly outperform passive tactile or vision-only baselines. Future work will explore bimanual manipulation, anthropomorphic hands, and flow-matching objectives to enhance policy reactivity.

References

[1] R. S. Johansson and G. Westling, “Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects,” *Experimental brain research*, vol. 56, no. 3, pp. 550–564, 1984.

[2] ——, “Signals in tactile afferents from the fingers eliciting adaptive motor responses during precision grip,” *Experimental brain research*, vol. 66, no. 1, pp. 141–154, 1987.

[3] A. Bicchi and V. Kumar, “Robotic grasping and contact: A review,” in *IEEE Int. Conf. on Robotics and Automation*, 2000.

[4] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, “Tactile sensing—from humans to humanoids,” *IEEE T-RO*, 2010.

[5] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters *et al.*, “An algorithmic perspective on imitation learning,” *Foundations and Trends® in Robotics*, 2018.

[6] 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,” 2025.

[7] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, “Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots,” in *RSS*, 2024.

[8] E. Helmut, L. Dziarski, N. Funk, B. Belousov, and J. Peters, “Learning force distribution estimation for the gelsight mini optical tactile sensor based on finite element analysis,” in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 8553–8560.

[9] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *IJRR*, 2024.

[10] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, “On the continuity of rotation representations in neural networks,” in *CVPR*, June 2019.