2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Tactile-Conditioned Diffusion Policy for Force-Aware Robotic Manipulation

Erik Helmut[1], Niklas Funk[1], Tim Schneider[1], Cristiana de Farias[1], Jan Peters[1,2,3,4,5]

***Abstract*— Contact-rich manipulation depends on applying the correct grasp forces throughout the manipulation task, especially when handling fragile or deformable objects. Most existing imitation learning approaches often treat visuotactile feedback only as an additional observation, leaving applied forces as an uncontrolled consequence of gripper commands. In this work, we present Force-Aware Robotic Manipulation (FARM), an imitation learning framework that integrates high-dimensional tactile data to infer tactile-conditioned force signals, which in turn define a matching force-based action space. We collect human demonstrations using a modified version of the hand-held Universal Manipulation Interface (UMI) gripper that integrates a GelSight Mini visual tactile sensor. For deploying the learned policies, we developed an actuated variant of the UMI gripper with geometry matching our hand-held version. During policy rollouts, the proposed FARM diffusion policy jointly predicts robot pose, grip width, and grip force. FARM outperforms several baselines across three tasks with distinct force requirements—high-force, low-force, and dynamic force adaptation—demonstrating the advantages of its two key components: leveraging 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 a fruit or an egg. In such cases, it is essential to employ the appropriate grasping force to minimize the risk of slippage or breakage. Tactile sensing has therefore emerged as a key component of forceful manipulation, enabling slip detection and inference of shear and normal forces [4]. Yet, despite its importance, effectively harnessing tactile sensing for direct force control in robots remains challenging.

One direction to address this challenge is imitation learning, which has emerged as an efficient approach to robotic manipulation by leveraging human demonstrations [5]. Recent work has integrated tactile feedback into these methods.

Corresponding author: Erik Helmut. Email: erik@robot-learning.de.
[1]TU Darmstadt, Institute for Intelligent Autonomous Systems, Computer Science Department [2]German Research Center for AI (DFKI), Research Department: Systems AI for Robot Learning [3]Hessian Center for Artificial Intelligence (hessian.AI) [4]Robotics Institute Germany (RIG) [5]TU Darmstadt, Centre for Cognitive Science

This work has been partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG) and project DEMETER (Grant no.: 01DR25003), 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).

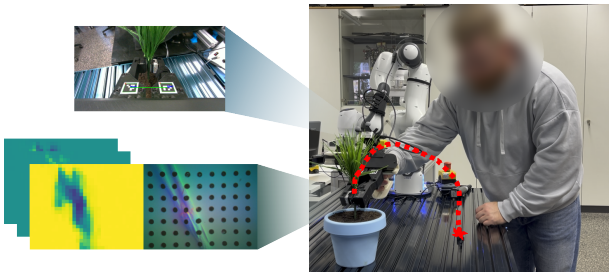


Fig. 1: Data-collection setup using the adapted hand-held UMI gripper. Right: An expert performs a task using the adapted hand-held UMI gripper. Top left: In-hand RGB camera view with ArUco markers for grip width measurement. Bottom left: GelSight Mini tactile image and corresponding FEATS force estimates, visualizing the contact interaction and force distributions during demonstration.

However, in most cases, tactile sensing is treated mainly as an additional observation modality, useful to get information for resolving visual occlusion or detecting contact state, but not as a signal that directly shapes the action space [6]. As a result, tactile feedback influences the policies only indirectly through its effect on the observation embedding, while the applied forces themselves remain an uncontrolled consequence of gripper commands. What remains largely missing is an imitation learning framework in which tactile feedback is not only perceived but also explicitly incorporated in the action space. Such a formulation would allow the policy to target and regulate the tactile interaction it intends to produce, rather than leaving contact forces as an uncontrolled side-effect of kinematic and gripper control.

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. For this purpose, we leverage the GelSight Mini [7], a high-resolution vision-based tactile sensor, together with Finite Element Analysis for Tactile Sensing (FEATS) [8] for estimating shear and normal contact forces. The demonstrations are collected using an adapted hand-held Universal Manipulation Interface (UMI) gripper [9], as shown in Fig. 1. Using the expert demonstrations, we train a diffusion policy, which we execute on a Franka Research 3 robot with an actuated variant of the UMI gripper. The *Actuated UMI* gripper, which we introduce in this work, has a matching geometry with our hand-held UMI gripper, and thus functions as a drop-in replacement without the need for retargeting. During execution, the gripper operates

under a dual-mode control scheme: it alternates between position control of the grip width and closed-loop force control, depending on the presence of contact. This approach ensures stable and reliable performance in both free-space and contact conditions. The main contributions of this work are summarized as follows:

- We present the FARM diffusion policy, which predicts robot pose, target grip width, and target grip force jointly, with forces represented both in the observation space and as an explicit action, yielding temporally consistent, force-aware action sequences.
- We design and build the *Actuated UMI* gripper as an open-source platform, enabling direct transfer of demonstrations collected with the adapted hand-held UMI gripper equipped with a GelSight Mini sensor.
- We analyze and ablate the effects of different control modes and tactile representations on task success rates across three real-world tasks: plant insertion, grape picking, and screw tightening.

II. Related Works

Tactile sensing is a key modality for advancing contact-rich robotic manipulation [4], [10], [11], complementing vision by providing information about forces [8], [12], [13], texture [14], [15], and slip [16], [17], [18]. Indeed, tactile feedback is increasingly being integrated into learning-based manipulation frameworks [19]. In the context of visuotactile imitation learning, prior work can broadly be divided into two categories: approaches that use the tactile feedback to enrich the observation space, and approaches that aim to exploit tactile information more directly in shaping the learned actions.

A. Incorporating Visuotactile Observations into Robotic Manipulation

Works such as 3D-ViTac [20], GelFusion [21], and TactileAloha [22] integrate tactile and visual signals into a unified latent representation, enabling policies to either overcome visual occlusion or to leverage contact information during manipulation. These, however, all rely on teleoperation for data collection. Alternatively, works such as Ablett et al. [23] rely on data collected through kinesthetic teaching. In this case, the kinesthetic data is converted back into robot actions through a tactile force matching objective that is active when replaying the trajectories using an impedance controller. This replay of the trajectories collected through kinesthetic teaching then yields the pose setpoints which are ultimately used for policy training. This idea conceptually aligns with DexForce [24], which leverages contact forces, measured on a robotic hand with F/T sensors during kinesthetic demonstrations. By converting the measured forces into force-informed position targets via an impedance controller, the robot can replay the demonstrations, yielding trajectories suitable for policy learning. While all these methods highlight the utility of leveraging tactile information for policy learning, tactile feedback remains an auxiliary observation signal rather than a modality that is directly exploited within the controller. Moreover, these methods also rely on demonstration data collection with the robot-in-the-loop either through kinesthetic teaching and subsequent force-informed trajectory replay [23], [24], or teleoperation [20], [21], [22]. Such strategies inherently couple the demonstrations to the specific robot embodiment, which can limit generalization across platforms. By contrast, we adopt a robot-free data collection paradigm using an adapted hand-held UMI gripper that records visuotactile observations, thereby mitigating the embodiment gap.

In parallel, recent works have also adopted a robot-free data collection paradigm through custom visuotactile manipulation interfaces. MimicTouch [25] builds on non-parametric imitation learning, where tactile and audio embeddings together with the robot end-effector pose are matched against a demonstration library to retrieve a nearest-neighbor-based action prediction. Online residual reinforcement learning is then used to adapt the policies learned from human demonstrations for robotic execution. ViTaMIn [26] exploits a customized Universal Manipulation Interface (UMI) gripper and proposes a multimodal representation learning strategy to obtain a tactile representation that captures essential contact properties, such as the object's in-hand pose and gripper's deformation. FreeTacMan [27] demonstrate that tactile features combined with visual observations improve success in contact-rich manipulation tasks compared to vision-only approaches. However, compared to this work, in all of these systems, the policy ultimately outputs joint or Cartesian commands, leaving tactile feedback as an indirect signal that is not directly exploited for representing the actions.

B. Tactile Sensing for Action Representation

Instead of passively conditioning the policy on tactile signals, some works use these signals to regulate actions, marking an important step toward force-aware control.

For instance, Liu et al. [28] proposed ForceMimic, a force-centric robot learning system that collects human demonstrations using a hand-held, robot-free device (ForceCapture) with an integrated 6-axis F/T sensor mounted between the gripping handle and the gripper. This setup enables direct recording of interaction forces with the environment. Using this data, they train a diffusion policy to predict wrench-pose trajectories and apply hybrid force-position control to fit the predicted force-position parameters.

Xu et al. [6] developed the TactAR teleoperation system to collect demonstration data with real-time visual tactile/force feedback. Building on this data, the authors propose a Reactive Diffusion Policy, where a latent diffusion policy predicts action chunks in latent space at low frequency, while the fast asymmetric tokenizer refines these latent actions at high frequency using real-time tactile feedback, effectively acting as a learned impedance controller. While their methods focus on controlling the forces exerted by the robot arm, our approach specifically controls the grip force, where grip force and gripper width are themselves actions predicted by the policy. By embedding forces directly

into the action representation, our method provides finer-grained control at the contact interface. Furthermore, while the reactive diffusion policy separates planning and reactive refinement into two subsystems, our single diffusion policy jointly predicts the robot pose, grasp width, and grasp force trajectories.

Closely related to our work, Adeniji et al. [29] introduce the Feel-the-Force framework, which also incorporates tactile sensing directly into the action space. Using human demonstrations collected with a tactile glove, their policy predicts gripper end-effector poses and grasping forces, which are then executed through a PD force controller. This approach, however, relies on a calibrated setup and a reset alignment between the human hand and robot gripper, as well as manual annotation of semantic keypoints to initialize scene representations. Moreover, their execution is constrained by binarized gripper states and requires the force controller to converge before the robot advances to the next action, which slows execution and limits adaptability. By contrast, our method learns directly from robot-embodied demonstrations through a hand-held gripper, avoiding cross-domain retargeting. Instead of a binary gripper state, we predict continuous grip width and target forces, enabling smoother and more precise control. A dual-mode controller scheme decides when to regulate grip width versus grip force, allowing execution to proceed without waiting for force convergence and enabling adaptation to real-time tactile feedback.

Together, these works demonstrate the promise of incorporating tactile sensing into the action space. Yet, they either do not control the grasp force at all or use a restrictive, synchronized control scheme, in which the agent cannot control the target gripper width. Our approach provides advances in this direction by directly coupling continuous grip force control with diffusion policy learning from robot-embodied tactile data, enabling a continuous, adaptive, and contact-aware manipulation.

III. Method

Here, we introduce Force-Aware Robotic Manipulation (FARM), an imitation learning framework to learn tactile-informed manipulation policies from demonstration data. To capture detailed contact interactions during manipulation, we integrate the GelSight Mini sensor into a custom-built robotic gripper, enabling high-resolution tactile feedback at the fingertip. From the sensor's raw tactile images, we extract estimates of the applied contact forces using Finite Element Analysis for Tactile Sensing (FEATS) [8]. This force information is integrated as both observation and action in a diffusion policy, which is trained to replicate human demonstrations not only in terms of gripper motion, but also by explicitly predicting and controlling the target grip force applied to the object. This framework is designed to generalize across tasks requiring either strong or delicate manipulation. The following sections describe the gripper hardware, data collection and processing pipeline, the design of the FARM diffusion policy, and the implementation of closed-loop force control for deployment on a real robot.

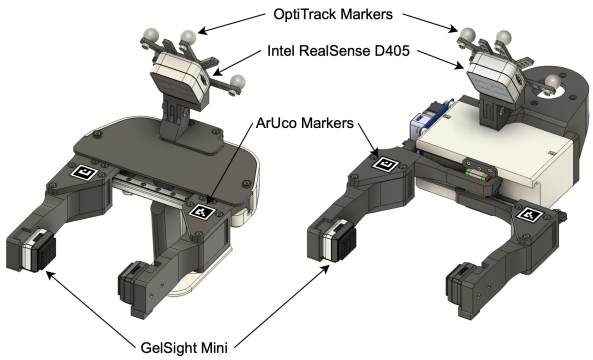


Fig. 2: Side-by-side comparison of the adapted UMI gripper (left) used for demonstration data collection and the Actuated UMI gripper (right) used for robotic deployment. Both feature an Intel RealSense D405 camera, a GelSight Mini tactile sensor (mounted on one fingertip), OptiTrack markers for motion tracking, and ArUco markers for grip width measurement. Sensor placement and overall geometry are matched to enable direct transfer of learned policies to the Actuated UMI.

A. Gripper Hardware

Our approach relies on two closely related grippers: an adapted hand-held Universal Manipulation Interface (UMI) gripper [9] for demonstration data collection, and a custom-built Actuated UMI gripper for robotic deployment (Fig. 2).

Our modified version of the UMI gripper replaces the original GoPro camera with an Intel RealSense D405, which provides in-hand RGB images and is tracked via OptiTrack markers for precise motion capture. The standard elastic TPU fingers are replaced with rigid fingers. One finger is fitted with a GelSight Mini sensor at its fingertip, and the other finger holds a shell of the GelSight Mini sensor with a matching gel pad but no electronics. In addition, we attach ArUco markers to each finger and use the in-hand Intel RealSense D405 camera to track their positions, enabling precise measurement of the grip width.

To deploy learned policies on a real robot, we developed the Actuated UMI gripper. Powered by a single DYNAMIXEL XL430-W250-T motor, this gripper uses a belt-driven mechanism to synchronously actuate both fingers. Two coil springs, one per finger, return the gripper to its open position when torque is released. The geometry mirrors the adapted hand-held UMI gripper. All sensors and markers, including the GelSight Mini, the Intel RealSense D405, the OptiTrack markers, and the ArUco markers, are positioned identically, enabling direct transfer of policies learned on demonstration data. The Actuated UMI supports multiple control modes, such as position, velocity, pulse width modulation, and operates at approximately $50\,\mathrm{Hz}$, making real-time force control feasible. We open-source all the design files and control software.

B. Data Collection

By using the adapted hand-held UMI gripper, we are able to both demonstrate the required motions and directly

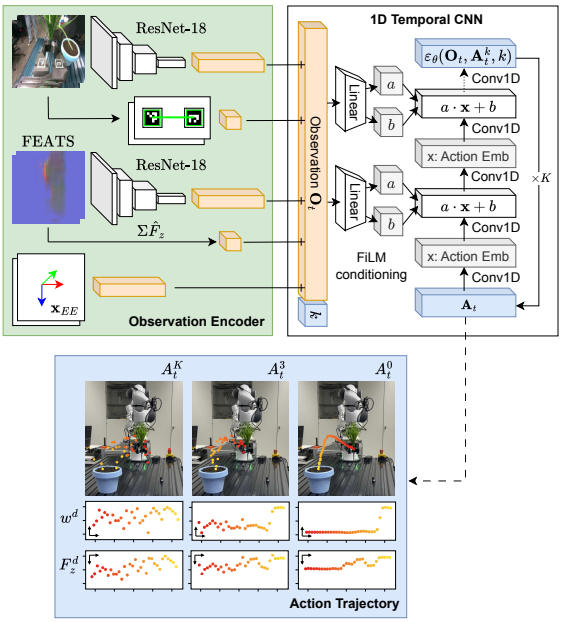


Fig. 3: FARM diffusion architecture. Visual, proprioceptive, and tactile observations are encoded and provided as input to a 1D temporal CNN with FiLM conditioning. The model predicts action trajectories including absolute end-effector pose, grip width, and grip force. This structure enables closed-loop force control of the gripper during manipulation.

apply the necessary contact forces for each task. Capturing both gripper kinematics and high-resolution tactile feedback during demonstrations enables us to record the force profiles essential for learning policies that require precise grip control, something not achievable with conventional teleoperation or kinesthetic teaching methods. Fig. 1 illustrates our demonstration setup along with the different sensor inputs.

During each demonstration, we record synchronized streams of all relevant sensor data: RGB images from the Intel RealSense D405 camera ($848 \times 480 \times 3$); grip width from the ArUco markers on the fingers, with positions determined by the in-hand Intel RealSense camera; gripper pose via OptiTrack, (the rigid body frame is defined by the marker constellation); GelSight Mini tactile images ($320 \times 240 \times 3$); force distribution estimates computed from each GelSight Mini image using the open-source FEATS model.

FEATS [8] is a learning-based method that infers spatial force distributions from GelSight Mini images. It is trained on labeled data generated via finite element analysis. FEATS thus outputs physically grounded estimates of the shear and normal forces. We note that although FEATS demonstrates strong performance, deviations in the predictions may occur due to underlying model assumptions or challenges in generalizing across sensors. Additionally, we represent the gripper pose through its 3D position and a 6D rotation representation [30]. All data is collected using Robot Operating System (ROS). After recording, we synchronize all sensor streams to the GelSight Mini images, which operates at 25 Hz and act as the reference clock. This ensures that every sample in the trajectory contains a complete set of sensor observations.

C. FARM Diffusion Policy

We build on the diffusion policy introduced by Chi et al. [31], using the 1D temporal Convolutional Neural Network (CNN) variant with Feature-wise Linear Modulation (FiLM) conditioning [32]. To incorporate tactile feedback into the diffusion policy in a physically meaningful and interpretable way, we extend the observation and action spaces beyond those used in prior work [31], [9]. Each policy input at time t includes:

1) **In-hand RGB image** from Intel RealSense D405 camera, downsampled to $96 \times 96 \times 3$ for computational efficiency without sacrificing the essential visual context.
2) **Grip width**, calculated as the Euclidean distance between the centers of the ArUco markers on the two fingers, measured in the Intel RealSense D405 image.
3) **Tactile feedback** is represented by a 3-channel image encoding of the force distributions extracted from the GelSight Mini tactile image using a pretrained and fixed FEATS model. The image channels are (1) the shear force in the x-direction, (2) the shear force in the y-direction, and (3) the normal force in the z-direction, with each channel independently normalized and the overall image scaled to a $96 \times 96 \times 3$ resolution. Additionally, the total normal force $\hat{F}_z$ is included as a scalar value, computed by integrating over the discretized normal force distribution, as it directly corresponds to the quantity regulated during closed-loop force control on the gripper.
4) **Gripper pose**, consisting of 3D position coordinates and 6D rotation feature representation [30] providing an continuous encoding of spatial orientation.

To process the visual inputs, we employ separate ResNet-18 encoders for the in-hand RGB image and the tactile force image, enabling specialized feature extraction for each modality. The diffusion policy is designed to predict action trajectories of the absolute target pose, target grip width w^d, and target absolute grip force F_z^d, each over a fixed prediction horizon of 32 and action execution horizon of 16. The observation horizon comprises two observations, which allows the policy to capture short-term context for decision-making (Fig. 3). By incorporating both force and grip width into the state and action space, these quantities are used for conditioning the model as well as being treated as output variables during the denoising process. This design ensures that the current grip force and grip width observations directly influence the predicted action trajectory, including the target force and target grip width at each step. As a result, the model's predictions align with current observations, mitigating the risk of implausible or unstable grasping behavior. This enables the policy to anticipate and regulate the force required for subsequent steps.

The diffusion policy is trained by minimizing the mean squared error between the predicted noise and the actual noise added to clean action samples. We use the implementation from LeRobot [33] with modifications to support our extended observation and action modalities.

D. Policy Deployment and Gripper Control

Including both target grip width and target force in the action space allows us to capture both positional and force-related aspects of manipulation. Target grip force is the relevant control variable during object contact because it enables closed-loop force control. However, target grip width is required to guide finger positioning during phases when there is no contact, such as when approaching, grasping, or releasing an object. Without explicit grip width actions, the policy would lack the means to open or close the gripper accurately outside the contact phase.

To deploy learned policies on the Actuated UMI gripper, we therefore implement a dual-mode control strategy that switches between grip width control and force control based on the current interaction phase. This strategy allows the robot to execute both pre-contact motions and in-contact, closed-loop force control. The diffusion policy outputs both a target grip width w^d and a target force $F_z^d < 0$ at each step. The controller monitors both the target force and the estimated contact force $\hat{F}_z$, computed from FEATS using the latest GelSight Mini image I_{GS}. If both the target and estimated force are below $-0.5\,\mathrm{N}$, the system assumes that the robot is in contact with the object and switches to force control. Otherwise, the target grip width w^d is directly sent to the internal PD controller of the DYNAMIXEL motor. The switching threshold of $-0.5\,\mathrm{N}$ was selected based on the noise characteristics of the FEATS model, ensuring that the controller only transitions to force control when actual contact is confidently detected. The controller computes the force error $e = \hat{F}_z - F_z^d$ and applies a PID controller to this error with anti-windup to ensure the integral term cannot accumulate beyond actuator limits. This value is added to the current grip width and sent as a position command to the internal PD controller of the DYNAMIXEL motor. The force control loop runs at $25\,\mathrm{Hz}$, synchronized with GelSight Mini image acquisition and FEATS prediction. During deployment, the diffusion policy runs at $7\,\mathrm{Hz}$, while the lower-level force and position controllers operate at higher rates to bridge the gap between high-level action selection and real-time motor actuation.

To further ensure seamless transfer from demonstration data to robot execution, two calibrations are performed. First, hand-eye calibration [34] aligns the OptiTrack world frame with the robot base. This transformation allows us to directly command end-effector positions to the robot in its own base frame during policy execution. Second, a linear mapping between grip width, measured as the ArUco marker distance, and motor position is estimated via least squares by slowly closing the Actuated UMI gripper and recording both quantities. This ensures that grip width values predicted by the policy can be accurately converted into motor commands during deployment.

E. Baselines

To evaluate the contribution of tactile feedback and force control in our FARM framework, we compare it against three baseline strategies: (1) Force-Aware, (2) Tactile-Aware, and (3) Vision-Only. Each baseline is also a diffusion policy and operates on a modified input representation and/or action space, while keeping the proprioceptive (end-effector pose, grip width) and visual (RGB image from Intel RealSense D405) inputs consistent with Section III-C.

1) Force-Aware Baseline: This baseline uses the total normal force estimated by FEATS from GelSight Mini images as the only tactile input, alongside vision, end-effector pose, and grip width. The action space includes target end-effector pose, target grip width, and target grip force, enabling closed-loop force control via the dual-mode controller described in Section III-D. This baseline follows a similar pipeline as FARM, but omits the full force distribution, which limits its ability to capture detailed contact configurations.

2) Tactile-Aware Baseline: The tactile-aware baseline incorporates raw GelSight Mini images ($96 \times 96 \times 3$) as input, processed by a separate ResNet-18 encoder, alongside vision, end-effector pose, and grip width. The action space includes target end-effector pose and target grip width but excludes target force, preventing explicit force control. The reason for the lack of explicit force control is that while tactile feedback from raw GelSight Mini images provides implicit contact awareness, the policy cannot actively regulate contact forces based on having only the raw tactile images available.

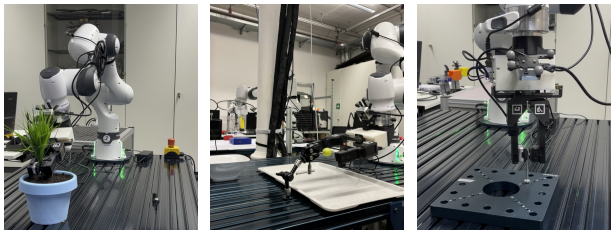
3) Vision-Only Baseline: The vision-only baseline omits tactile feedback entirely, relying on vision, end-effector pose, and a binary gripper state (open/closed) derived from thresholding the grip width in demonstration data. The action space includes the target end-effector pose and a binary open/close command. During inference, the gripper moves to fully open or closed positions using the DYNAMIXEL motor's internal PD controller.

IV. Experiments & Results

In this section, we evaluate our FARM framework that combines physically-grounded high-dimensional tactile observations on the input level together with force-based grip control. To investigate the contribution of both components, we compare the approach with the previously mentioned baselines, thereby assessing the role of tactile sensing and its representation for both the observation and action levels. We structure the experiments along three main questions: (1) *How important is the explicit incorporation of force signals for task success?* (2) *How does the high-dimensional (image-based) force-distribution information impact task success?* and (3) *How do these approaches compare in the force domain for the time-dependent task of screw tightening?*

A. Experiment Setup

All experiments are conducted using a real Franka Research 3 robot equipped with the Actuated UMI gripper (see Section III-A). The robot is controlled using the Cartesian position impedance controller from the franky control library [35]. Across all strategies, actuation of the Actuated UMI gripper relies on the internal PD position controller of its DYNAMIXEL motor.



(a) (b) (c)

Fig. 4: Experiment setup for the three tasks: (a) plant insertion, showing the Franka Research 3 robot planting a plastic plant into a soil-filled pot; (b) grape picking, depicting the delicate grasping and detachment of a grape from a toothpick; (c) screw tightening, illustrating the robot using an Allen key to tighten a screw.

For all tasks, we collected 30 demonstrations with our hand-held UMI gripper (see Section III-B), all performed by the same expert. We train the diffusion policies using Denoising Diffusion Probabilistic Models (DDPM) with 100 denoising steps over 60000 iterations. For inference, we adopt Denoising Diffusion Implicit Models (DDIM) with 10 steps to reduce the number of sampling iterations [31].

The experiments involve three tasks, shown in Fig. 4, each testing different aspects of tactile-informed manipulation:

Plant Insertion Task. The task requires grasping a plastic plant from a fixed holder, inserting it into a flower pot filled with moistened soil, and finally releasing it. This task requires a high grip force to ensure stable insertion without losing grasp. A rollout is considered successful if the plant remains upright in the soil without tilting or touching the pot rim. To standardize the initial conditions across all approaches, we constrain the rollout until grasp detection.

Grape Picking Task. This task consists of grasping a grape from a toothpick without crushing or letting it slip from the robot's grip. Successful execution here consists of grasping the grape, delicately removing it from the toothpick, and finally, placing it into a bowl. A rollout is successful if the grape is placed intact and visually undamaged into the bowl.

Screw Tightening Task. This task requires grasping an Allen key inserted in a screw at a fixed position, rotating it, and opening the gripper upon having tightened the screw. To introduce additional variability into the task, we consider two relative orientations between the Allen key and the screw, resulting in variation in the final tightened position. Success is achieved if the screw is noticeably tight. This task relies on tactile feedback to detect tightness, as vision alone cannot determine the screw's state. For safety, rollouts are terminated if the end-effector torque exceeds a threshold.

B. Results

1) How important is the explicit incorporation of force signals for task success?: To evaluate the importance of explicit force signals over raw tactile images or vision-only inputs, we compare the force-aware, tactile-aware, and vision-only baselines. Success rates are assessed over 20 rollouts per task, with results summarized in Fig. 5.

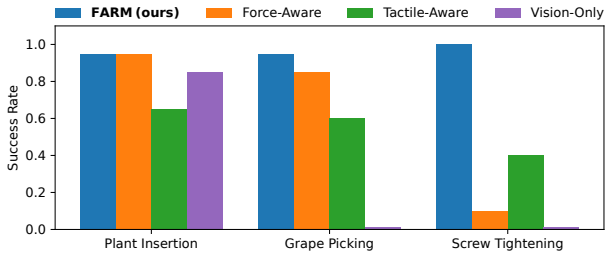


Fig. 5: Success rates for the plant insertion, grape picking, and screw tightening tasks, comparing the FARM framework against force-aware, tactile-aware, and vision-only baselines, evaluated over 20 rollouts per task.

In the plant insertion task, the force-aware baseline achieves a 95% success rate, surpassing tactile-aware, with 65% success rate, and the vision-only baseline with 85% success. The explicit force control in the force-aware baseline ensures sufficient grip to prevent slippage during plant transport and insertion into soil. Tactile-aware baseline, lacking direct force modulation, frequently applies insufficient grasp force, leading to slippage or incomplete planting. The vision-only baseline performs moderately well due to its binary gripper command by applying high grasping forces, but its lack of tactile feedback limits its ability to confirm stable grasps and successful plant insertion, resulting in occasional failures by opening the gripper too early.

In the grape picking task, the force-aware baseline achieves 85% success rate, compared to the tactile-aware baseline's 60% success and vision-only 0% success. The force-aware baseline's closed-loop force control applies precise force to avoid crushing the grape while maintaining a secure grip. With the tactile-aware baseline's passive use of raw tactile images, it often results in weak grips, causing slippage. The vision-only baseline fails entirely, as its coarse binary gripper command consistently crushes the grape.

Finally, in the screw-tightening task, where tactile feedback is essential for detecting screw tightness, the force-aware baseline achieves 10% success rate and the tactile-aware baseline 40%. Both surpass the vision-only baseline, which once again fails entirely. In this case, due to a lack of tactile feedback needed to sense screw resistance, it consistently opens the gripper prematurely. The scalar force signal available to the force-aware baseline enables partial detection of this resistance but is insufficient for maintaining proper Allen key alignment, often causing it to slip out of the screw head. In contrast, the tactile-aware baseline leverages higher-dimensional information to outperform the force-aware baseline. However, this baseline also frequently opens the gripper too early, missing tightness cues due to its lack of explicit force measurements.

These results highlight that explicit force incorporation, as in the force-aware baseline, improves task success in manipulation tasks that require consistent static high or low forces, such as plant insertion and grape picking. However, its limited performance in the screw tightening task indicates that scalar force signals alone may be insufficient for tasks

requiring precise contact state detection, suggesting the need for richer force representations. To address this, in the next sections, we investigate how incorporating the full force distribution information impacts task success.

2) How does the high-dimensional (image-based) force-distribution information impact task success?: Building on the established importance of explicit force signals, we now examine how high-dimensional, image-based force-distribution information, as used in the FARM framework, enhances task success compared to scalar force signals, raw tactile images, and no tactile feedback. FARM encodes shear and normal forces as images (Fig. 3), providing richer contact information than the scalar normal force in the force-aware baseline or the raw tactile image in the tactile-aware baseline.

As seen in Fig. 5, in the plant insertion and grape picking tasks, FARM achieves success rates of 95% for both tasks, which is comparable to the force-aware baseline with 95% and 85% respectively. These results indicate that scalar force signals suffice for tasks with consistent static force demands. The screw tightening task best highlights the advantages of force distributions over scalar force signals or raw tactile images, with FARM achieving a 100% success rate compared to force-aware 10%, tactile-aware 40%, and vision-only 0% baselines. FARM's tactile force images capture shear and normal force interactions, enabling both a more precise downward pressure to keep the Allen key engaged and the accurate detection of screw tightness. In comparison, the force-aware baseline relies only on a scalar normal force signal, struggling to maintain proper Allan key alignment and often lifting it out of the screw head. The tactile-aware, while better than the force-aware policy, still fails to reliably interpret the complex contact dynamics from raw tactile images, leading to premature gripper opening. The vision-only baseline lacks the needed tactile feedback for this task.

These results show that high-dimensional force-distribution information significantly enhances task success, particularly in tasks requiring nuanced force interactions. By capturing both shear and normal forces, FARM provides superior control and contact state awareness across all tasks compared to scalar force signals or raw tactile images. This advantage is most pronounced in tasks where temporal and multi-dimensional force dynamics are critical.

3) How do these approaches compare in the force domain for the time-dependent task of screw tightening?: To answer this question, we compare FARM in the force domain against the force-aware, tactile-aware, and vision-only baselines in the time-dependent screw tightening task. To quantify how closely the applied forces from each approach match those of the demonstrations, we use the Wasserstein-1 distance W_1 [36]. Since trajectories differ in duration, simply pooling all samples would bias the statistics toward longer trajectories. For that reason, we use an equal-mass weighting scheme, ensuring that each demonstration or policy rollout contributes equally regardless of trajectory length. For trajectory m with n_m samples, each sample receives the weight $w_{k,m} = 1/M \cdot n_m$, where M is the number of trajectories.

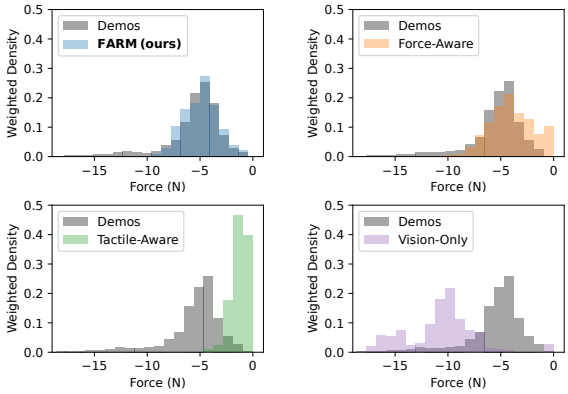


Fig. 6: Weighted empirical force distributions for the screw tightening task, comparing demonstrations and rollouts for the FARM framework against force-aware, tactile-aware, and vision-only baselines.

The W_1 distance is then defined as

$$W_1(u, v) = \inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| d\pi(x, y),$$

for weighted empirical force distributions of demonstrations u and policy rollouts v. The W_1 value corresponds to the average amount the force distributions from one set would need to be shifted to match the other, expressed in the same unit as the force (N). Smaller values, therefore, indicate that the forces measured during rollouts are more similar to those of the demonstrations, while larger values indicate greater differences in either magnitude or distributional shape.

As shown in Fig. 6, FARM's weighted empirical force distributions closely align with those of the demonstrations. This is also represented by the Wasserstein-1 distance of $0.7538\,\mathrm{N}$. The force-aware baseline follows with a distance of $1.6580\,\mathrm{N}$, showing reasonable, but less precise force matching. Although the force-aware baseline was able to produce similar forces, it nevertheless failed in the task success because it lacked the necessary contact state information from just a single force value. The tactile-aware baseline, with a Wasserstein-1 distance of $4.3801\,\mathrm{N}$, tends to apply insufficient forces, while the vision-only baseline, at $5.0515\,\mathrm{N}$, consistently applies excessive forces, leading to significant deviations from the demonstration force profiles. These results show that incorporating force distributions into the observation embedding enables FARM to outperform the baselines in the force domain. The larger W_1 distances for force-aware, tactile-aware, and vision-only baselines reflect their limitations in capturing the complex, time-varying force profiles required for precise screw tightening.

V. Conclusion

This work investigated how tactile sensing can be incorporated not only as an observation modality but also directly into the action space of imitation learning policies for contact-rich manipulation. We thus introduced FARM, a visuotactile-conditioned diffusion policy that predicts target

grip width and target grip force jointly with the robot pose. Within FARM, the tactile-conditioned contact information is used both as an observation and as a basis for generating force-based action sequences. Demonstrations were collected using an adapted hand-held UMI gripper equipped with a GelSight Mini sensor at one fingertip, with normal and shear force distributions estimated via the FEATS model. To enable policy transfer, we developed the *Actuated UMI* gripper, whose geometry and kinematics match the hand-held UMI gripper. We executed gripper actions through a dual-mode controller that switches between grip width position control and closed-loop force control. The real-world experimental results and comparison with several baselines show the effectiveness of our proposed method and underline the importance of both leveraging a force-based action space, as well as physically-grounded, high-dimensional tactile information as an observation modality. In addition to improved reliability and success rates, our proposed method also aligns more closely with the demonstration data w.r.t. the applied forces. While FARM has shown promising results on three diverse tasks, including grasping objects of varying stiffness and maintaining a stable grip during screw tightening, we aim to extend it to more unstructured, in-the-wild settings. Future work should explore generalizing the method to bimanual manipulation, anthropomorphic hands, and flow-matching objectives to enhance the 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] GelSight, Inc., "Gelsight mini - datasheet," https://www.gelsight.com/wp-content/uploads/2023/01/GelSight_Datasheet_GSMini_12.20.22.pdf, accessed: 2025-06-08.

[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, 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.

[10] M. R. Cutkosky and J. M. Hyde, "Manipulation control with dynamic tactile sensing," in *6th international symposium on robotics research*, 1993.

[11] W. Mandil, V. Rajendran, K. Nazari, and A. Ghalamzan-Esfahani, "Tactile-sensing technologies: Trends, challenges and outlook in agri-food manipulation," *Sensors*, 2023.

[12] N. Funk, P. O. Müller, B. Belousov, A. Savchenko, R. Findeisen, and J. Peters, "High-resolution pixelwise contact area and normal force estimation for the gelsight mini visuotactile sensor using neural networks," in *Embracing Contacts-Workshop at ICRA 2023*, 2023.

[13] C. Sferrazza, A. Wahlsten, C. Trueeb, and R. D'Andrea, "Ground truth force distribution for learning-based tactile sensing: A finite element approach," *IEEE Access*, 2019.

[14] A. Böhm, T. Schneider, B. Belousov, A. Kshirsagar, L. Lin, K. Doerschner, K. Drewing, C. A. Rothkopf, and J. Peters, "What matters for active texture recognition with vision-based tactile sensors," in *ICRA*. IEEE, 2024.

[15] L. R. Manfredi, H. P. Saal, K. J. Brown, M. C. Zielinski, J. F. Dammann III, V. S. Polashock, and S. J. Bensmaia, "Natural scenes in tactile texture," *Journal of neurophysiology*, vol. 111, no. 9, pp. 1792–1802, 2014.

[16] W. Chen, H. Khamis, I. Birznieks, N. F. Lepora, and S. J. Redmond, "Tactile sensors for friction estimation and incipient slip detection—toward dexterous robotic manipulation: A review," *IEEE Sensors Journal*, 2018.

[17] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a gelsight tactile sensor," in *ICRA*. IEEE, 2015.

[18] N. Funk, E. Helmut, G. Chalvatzaki, R. Calandra, and J. Peters, "Evetac: An event-based optical tactile sensor for robotic manipulation," *IEEE T-RO*, 2024.

[19] N. Funk, C. Chen, T. Schneider, G. Chalvatzaki, R. Calandra, and J. Peters, "On the importance of tactile sensing for imitation learning: A case study on robotic match lighting," 2025. [Online]. Available: https://arxiv.org/abs/2504.13618

[20] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d vitac:learning fine-grained manipulation with visuo-tactile sensing," in *CoRL*, 2024.

[21] S. Jiang, S. Zhao, Y. Fan, and P. Yin, "Gelfusion: Enhancing robotic manipulation under visual constraints via visuotactile fusion," 2025. [Online]. Available: https://arxiv.org/abs/2505.07455

[22] N. Gu, K. Kosuge, and M. Hayashibe, "Tactilealoha: Learning bimanual manipulation with tactile sensing," *IEEE RA-L*, 2025.

[23] T. Ablett, O. Limoyo, A. Sigal, A. Jilani, J. Kelly, K. Siddiqi, F. Hogan, and G. Dudek, "Multimodal and force-matched imitation learning with a see-through visuotactile sensor," *IEEE T-RO*, 2025.

[24] C. Chen, Z. Yu, H. Choi, M. Cutkosky, and J. Bohg, "Dexforce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2501.10356

[25] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao, "Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2310.16917

[26] F. Liu, C. Li, Y. Qin, A. Shaw, J. Xu, P. Abbeel, and R. Chen, "Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface," 2025. [Online]. Available: https://arxiv.org/abs/2504.06156

[27] L. Wu, C. Yu, J. Ren, L. Chen, R. Huang, G. Gu, and H. Li, "Freetacman: Robot-free visuo-tactile data collection system for contact-rich manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2506.01941

[28] 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 *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[29] A. Adeniji, Z. Chen, V. Liu, V. Pattabiraman, R. Bhirangi, S. Haldar, P. Abbeel, and L. Pinto, "Feel the force: Contact-driven learning from humans," 2025. [Online]. Available: https://arxiv.org/abs/2506.01944

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

[31] 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.

[32] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," 2017. [Online]. Available: https://arxiv.org/abs/1709.07871

[33] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec *et al.*, "Lerobot: State-of-the-art machine learning for real-world robotics in pytorch," https://github.com/huggingface/lerobot, 2024.

[34] R. Y. Tsai, R. K. Lenz *et al.*, "A new technique for fully autonomous and efficient 3 d robotics hand/eye calibration," *IEEE T-RO*, 1989.

[35] T. Schneider, "franky: High-Level Control Library for Franka Robots." [Online]. Available: https://github.com/TimSchneider42/franky

[36] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2021, vol. 58.