# FILIC: Dual-Loop Force-Guided Imitation Learning with Impedance Torque Control for Contact-Rich Manipulation Tasks

Haizhou Ge $^{1*}$ , Yufei Jia $^{1*}$ , Zheng Li $^{2*}$ , Yue Li $^{3*}$ , Zhixing Chen $^{1*}$ , Ruqi Huang $^{1\dagger}$ , Guyue Zhou $^{1\dagger}$ 

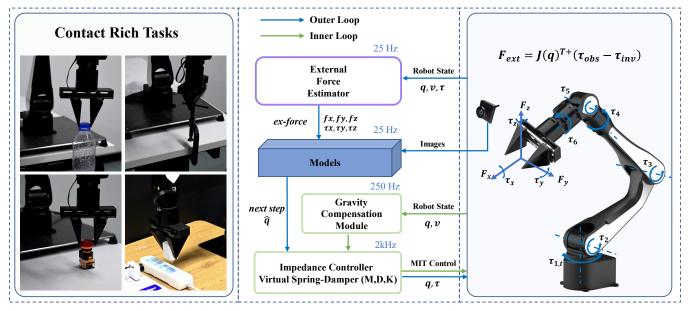


Fig. 1: **FILIC**. (**Left**) Examples of contact-rich manipulation tasks. (**Middle**) Dual-loop architecture: an outer-loop imitation learning model processes multimodal sensory inputs (estimated external force from an estimator and visual images from cameras) to predict the next action at 25 Hz; a gravity compensation module ensures accurate dynamics estimation at 250 Hz; the inner-loop impedance controller runs at 2 kHz for compliant torque control. (**Right**) Illustration of external forces estimation in joint space which are computed from joint torques using the robot's Jacobian.

Abstract—Contact-rich manipulation is crucial for robots to perform tasks requiring precise force control, such as insertion, assembly, and in-hand manipulation. However, most imitation learning (IL) policies remain position-centric and lack explicit force awareness, and adding force/torque sensors to collaborative robot arms is often costly and requires additional hardware design. To overcome these issues, we propose FILIC, a Forceguided Imitation Learning framework with impedance torque control. FILIC integrates a Transformer-based IL policy with an impedance controller in a dual-loop structure, enabling compliant force-informed, force-executed manipulation. For robots without force/torque sensors, we introduce a cost-effective end-effector force estimator using joint torque measurements through analytical Jacobian-based inversion while compensating with model-predicted torques from a digital twin. We also design complementary force feedback frameworks via handheld haptics and VR visualization to improve demonstration quality. Experiments show that FILIC significantly outperforms visiononly and joint-torque-based methods, achieving safer, more

compliant, and adaptable contact-rich manipulation. Our code can be found in https://github.com/TATP-233/FILIC.

#### I. Introduction

Contact-rich manipulation, where precise force and contact control are required, is essential for robots to interact effectively with objects and environments. It underpins tasks such as insertion [1], assembly [2], and in-hand manipulation [3]. Mastering contact-rich manipulation is crucial for achieving human-like dexterity and expanding robots' capabilities in complex, unstructured settings. Among various approaches, imitation learning (IL) offers significant potential in this context, as it enables robots to acquire complex manipulation skills from human demonstrations, facilitating efficient adaptation to diverse contact-rich tasks [4].

Despite recent progress, contact-rich manipulation still faces several key challenges. First, existing IL methods typically output joint positions or 6-DOF poses, which can generate excessive internal forces in contact-rich tasks, potentially damaging the robot or manipulated objects [5], [6]. Second, end-effector force provides richer and more intuitive contact information, which is crucial for contact-rich manipulation; however, robotic arms equipped with

<sup>\*</sup>Equal contribution; †Corresponding Author.

<sup>&</sup>lt;sup>1</sup>Tsinghua University. {ghz23,jyf23,chenzx24}@mails.tsinghua.edu.cn, ruqihuang@sz.tsinghua.edu.cn, zhouguyue@air.tsinghua.edu.cn

<sup>&</sup>lt;sup>2</sup>The Hong Kong University of Science and Technology (Guangzhou). zli514@connect.hkust-gz.edu.cn

<sup>&</sup>lt;sup>3</sup>DISCOVER Robotics. lue@discover-robotics.com

wrist force/torque sensors are scarce and often prohibitively expensive [7]. Third, collecting force data for contact-rich manipulation remains challenging, as humans lack the direct end-effector force perception that a robot requires, making purely visual demonstrations insufficient for capturing effective force information.

Several methods have been proposed to address these challenges, but each comes with limitations. Some IL algorithms incorporate end-effector force as an input to the model for training [8], [9]; however, they typically treat it merely as sensor information and perform only position control, without achieving true compliant manipulation, and often require expensive force-sensing robots. Other works attempt to use joint torque as input in the absence of end-effector force sensors, yet joint torque does not provide the same rich and intuitive contact information as end-effector force [1]. Force data collection also remains challenging: in kinesthetic demonstrations, human-applied external forces interfere with measurements, while in position-mapping demonstrations (e.g., VR or master-slave setups), operators lack force feedback and rely only on visual cues, making it difficult to capture effective end-effector force information [10].

To address these issues, we present **FILIC**, a Force-guided Imitation Learning framework with Impedance torque Control. First, we propose a novel dual-loop framework that integrates IL with impedance control. The Tranformer-based IL policy serves as the outer loop, taking visual images, end-effector pose and force as observations to predict target poses, which are then fed into the inner-loop impedance controller to generate torque commands, forming a forcein, force-out dual-loop structure that enables compliant manipulation. Second, for scenarios without end-effector force/torque sensors, we develop a cost-effective end-effector force estimator by leveraging the mapping between the robot's end-effector wrench and joint torques, based on the Jacobian matrix and a high-fidelity synchronized digital twin. Finally, to address challenges in data collection, we design two complementary force feedback frameworks: haptic feedback via handheld controllers and visual feedback in VR. Experimental results for insertion tasks in both simulation and real-world environments demonstrate that our approach significantly outperforms purely vision-based methods and those relying on joint torque alone.

This work makes the following contributions:

- A dual-loop framework combining an outer-loop Transformer-based imitation policy and an inner-loop impedance torque controller for compliant contact-rich manipulation with multimodal visual and force inputs.
- A low-cost, sensorless method for estimating endeffector wrench using a synchronous digital twin, with an open-source implementation for accessibility.
- Two complementary force-feedback data collection frameworks: haptic vibration via a handheld controller and AR-based force vector visualization in VR.

#### II. RELATED WORK

## A. Imitation Learning

Imitation learning (IL) trains policies from demonstrations, avoiding brittle reward engineering yet suffering from distribution shift in vanilla behavior cloning. Recent sequence or chunked action predictors improve temporal coherence and robustness [5], while diffusion-based policies formulate action generation as conditional denoising to capture multimodal futures [6]. Beyond action parameterization, data efficiency and generalization are advanced by 3D point-cloud conditioned diffusion policies (DP3) [11] and vision-language-action (VLA) models that inject largescale semantic priors [12]. Open-world variants ( $\pi 0$ ,  $\pi 0.5$ ) pair pre-trained VLMs with flow/matching action experts for broader task transfer [13], [14], and purely synthetic large-scale pretraining (GraspVLA) shows competitive zeroshot sim-to-real transfer [15]. Efficiency and deployability are pushed by compact architectures (SmolVLA) [16], lowlatency asynchronous chunk execution (RTC) [17], and stabilization techniques such as Knowledge Insulating [18]. Collectively, these trends point toward unified semantic reasoning plus temporally coherent control [19], yet current IL/VLA pipelines are still largely vision- and positioncentric, underutilizing rich physical interaction signals.

# B. Force-Aware Robot Control and Learning

Classical model-based strategies (impedance, admittance, direct force control) provide principled mappings between motion and interaction forces for safe, compliant behavior [20]–[22], and support stability across contact transitions in locomotion [23]. Yet in unstructured environments modeling errors, hybrid contact dynamics, and reality gaps limit purely analytical designs [24]. DRL plus domain randomization has produced agile, robust locomotion [25], leveraging dense proprioception and shaped rewards, but manipulation adds proactive environmental change [26], harder visual sim-to-real transfer, multimodal sensing demands [27], and sparse reward challenges [28]. Augmenting perception with large VLMs or distilled demonstrations [29] improves scene understanding but alone does not yield force-sensitive compliance.

Recent integration efforts target explicit force awareness. Diffusion policies over 6D wrenches achieve precise contact reasoning but often omit vision [30]. Variable/admittance impedance RL incorporates force—torque for high-precision assembly while still partially decoupling vision [31], [32]. Hierarchical or residual formulations extend the action space with stiffness/damping parameters [33], [34]. Stiffness inference from position demonstrations (SCAPE) reduces expert load yet remains position-centric and underuses rich force signals [35]–[37]. Parallel advances in sensorless force estimation leverage joint torque mappings as a low-cost proxy for contact without wrist F/T hardware, motivating torque-conditioned VLAs. Systematic studies of torque fusion and predictive torque modeling highlight gains from late fusion and auxiliary future-torque prediction for grounded internal

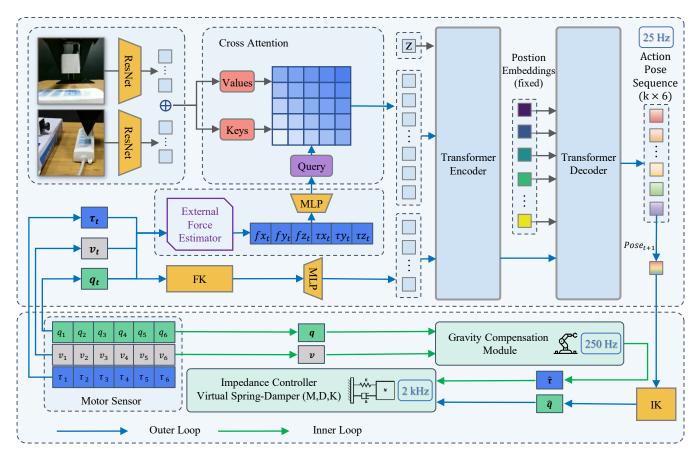


Fig. 2: **Detailed architecture of FILIC. Outer loop:** The external force estimator provides interaction forces, which are fused with forward kinematics (FK) outputs and further processed via MLP layers. Visual observations are encoded by dual ResNet backbones and integrated with estimated force embeddings through a cross-attention module. These multimodal features are passed into a standard Transformer architecture (e.g., ACT [5]) to generate pose sequences at 25 Hz. **Inner loop:** The predicted action pose is tracked through inverse kinematics (IK), while the inner control loop employs an impedance torque controller running at 2 kHz, supported by gravity compensation at 250 Hz.

representations [1]. Overall, the field is converging toward policies that couple semantic visual reasoning with learned, torque/force-aware control for compliant, contact-rich manipulation.

## III. PRELIMINARIES

## A. Imitation Learning

Imitation Learning (IL) is a paradigm in which an agent learns to perform tasks by observing and mimicking expert demonstrations. Formally, given a dataset of state-action pairs  $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^N$  collected from an expert policy  $\pi^*$ , the goal of IL is to learn a policy  $\pi_\theta$  that minimizes the discrepancy between  $\pi_\theta(a|s)$  and  $\pi^*(a|s)$ . Common approaches include Behavioral Cloning (BC), which treats IL as a supervised learning problem, and Inverse Reinforcement Learning (IRL), which infers a reward function underlying the expert's behavior. IL is particularly valuable in settings where designing reward functions is difficult or when high-quality demonstrations are available.

# B. Dynamics and Wrench Propagation in Manipulators

The dynamical model of a robotic manipulator is foundational for the analysis and control of interaction tasks. Given joint position q, velocity  $\dot{q}$ , and acceleration vectors  $\ddot{q}$ , the equations of motion are governed as follows:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + q(q) = \tau + \tau_{ext} \tag{1}$$

where M(q),  $C(q,\dot{q})$ , and g(q) are the inertia, Coriolis, and gravity terms, and  $\tau$  is the applied joint torques. In the presence of external contact, the joint external torque  $\tau_{ext}$  originates from a wrench  $F_{ext}$  applied at the end-effector.

A fundamental mapping arises from the kinematics of constrained motion: the effective wrench  $F_{ext}$  generated at the end-effector is algebraically related to the joint torques through the Jacobian transpose matrix  $J(q)^T$ :

$$\tau_{ext} = J(q)^T F_{ext} \tag{2}$$

This mapping is critical as it enables estimating external end-effector forces from joint current-derived torques without a dedicated force sensor. These inferred contact forces offer more intuitive physical information than raw

joint torques, greatly improving a model's understanding of interaction force space.

## C. Impedance Control

In robotic manipulation under unstructured environments, impedance control regulates the dynamic interaction between motion and external forces, which is typically modeled as a mass-spring-damper system, and can be implemented in either Cartesian or joint space. The desired joint torque  $\tau^*$  is computed as:

$$\tau^* = M(\ddot{q}_d - \ddot{q}) + D(\dot{q}_d - \dot{q}) + K(q_d - q) + \tau_{ff}$$
 (3)

where  $q_d$ ,  $\dot{q}_d$ ,  $\ddot{q}_d$  and q,  $\dot{q}$ ,  $\ddot{q}$  are the desired and measured joint angles, velocities, and accelerations, M, D, K are the joint-space inertia, damping, and stiffness matrices, and  $\tau_{ff}$  is the feedforward compensation torque.

During real-world deployment, acceleration measurements often suffer from significant errors, which can severely degrade control performance. As a result, the mass term is typically omitted, leading to a simplified impedance control formulation. Notably, such a impedance control can be realized through the MIT control mode of the actuators:

$$\tau_{mit}^* = K_d(\dot{q}_d - \dot{q}) + K_p(q_d - q) + \tau_{ff}$$
 (4)

where  $au_{mit}^*$  denotes the reference torque, while  $K_p$  and  $K_d$  denote the position and velocity gains, respectively.

#### IV. METHOD

## A. Dual-Loop Imitation-Impedance Control

In this work, to achieve imitation-learning-based compliant manipulation, we design a dual-loop control architecture that integrates a high-level imitation learning policy with a low-level impedance controller, as illustrated in Fig. 2. The outer loop employs imitation learning to generate adaptive task-level motion commands from multimodal observations, while the inner loop enforces stable and compliant execution through impedance control at the torque level.

Outer-loop imitation learning policy. We employ a Transformer-based model as the outer-loop imitation learning policy as shown by the blue flow in Fig. 1 and Fig. 2. To improve perception in contact-rich tasks, we incorporate a cross-attention mechanism that fuses force and visual inputs. The estimated force vector is encoded via an MLP as the query, while RGB observations are processed by two ResNets to provide the keys and values. The cross-attention module then aligns these modalities, allowing the policy to selectively attend to visual features most relevant to the contact dynamics, thereby capturing correlations between appearance and force interactions.

For state estimation, joint positions  $[q_1,...,q_n]_t$ , velocities  $[v_1,...,v_n]_t$ , and torques  $[\tau_1,...,\tau_n]_t$  are obtained from the robot's internal sensors and converted via forward kinematics and dynamics into the end-effector Cartesian pose and the corresponding external wrench  $F_{ext} = [fx_t, fy_t, fz_t, \tau x_t, \tau y_t, \tau z_t]$ . This proprioceptive information is synchronized with RGB images from two external cameras, all down-sampled to 25 Hz for efficient processing.

Each visual stream is encoded using a ResNet backbone, while the force vector is mapped through a multilayer perceptron (MLP). The resulting visual and force features are then fused via a cross-attention module. This fused representation is subsequently concatenated with two additional components: (1) the end-effector pose features, also projected through an MLP, and (2) a latent style variable Z, which is derived from encoding the action pose sequence and end-effector pose. Notably, Z is omitted during inference [5].

The Transformer-based model network outputs a sequence of target 6-DOF end-effector poses at 25 Hz, following the action chunking with temporal ensemble processing to enhance robustness and smoothness inspired by Action Chunking with Transformers (ACT) [5]. Except for the perception input module, the detailed architecture of the other modules follows ACT. These high-level references are passed to the inner-loop impedance controller, which translates them into compliant torque-level commands, thereby ensuring both task fidelity and safe interaction.

**Inner-loop impedance controller.** As depicted by the green flow in Fig. 1 and Fig. 2, to ensure robust and stable control on the physical platform, the impedance model is simplified to exclude the inertial component, since acceleration measurements are highly noise-prone and can severely degrade performance. The resulting formulation retains only the spring—damper terms, which balance compliance with stability.

At each control cycle, the imitation learning policy provides a 6-DOF Cartesian target pose  $pose_{t+1}$  at 25 Hz. This target is mapped into desired joint angles  $\hat{q}$  through real-time inverse kinematics. In parallel, joint positions q and velocities v are acquired from the robot's internal sensors at a higher rate of 250 Hz. These high-frequency proprioceptive signals are integrated into a joint-space impedance control law that computes torque commands  $\tau_{cmd}$  at 2 kHz formulated as:

$$\tau_{cmd} = B(\hat{v} - v) + K(\hat{q} - q) + \hat{\tau} \tag{5}$$

where K and B are the preset stiffness and damping coefficients,  $\hat{v}$  is typically set to 0, and  $\hat{\tau}$  is the feedforward gravity compensation torque, computed at 250 Hz. According to Equation (4) and (5),  $\hat{v}, \hat{q}, \hat{\tau}$  can be directly provided as inputs to the motor control function in MIT mode, from which the reference torque to be executed is computed.

In summary, the proposed hierarchical structure uniquely leverages end-effector forces as inputs for imitation learning and controls joint torques as outputs through impedance control, achieving robust, compliant, and demonstration-informed force control in contact-rich environments.

#### B. External Force Estimation via Digital Twin

**Jacobian-based external force estimation.** To exploit the intuitive and decoupled information provided by end-effector forces while overcoming the lack of such sensors in most affordable commercial robotic arms, we reconstruct the 6-dimensional end-effector wrench by mapping joint torques through the inverse relationship defined by the Jacobian, as presented the dark purple modules in Fig. 1 and Fig. 2.

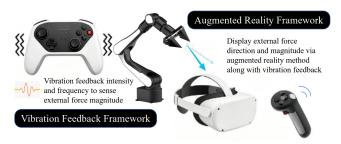


Fig. 3: Two frameworks of demonstration with force feedback. Wireless game controllers control robots in Cartesian space via joysticks and triggers, providing vibration-based force feedback, whereas VR devices track the operator's pose and render AR arrows on end-effectors to indicate force direction and magnitude.

The estimated force is then employed as perceptual input for imitation learning. Based on the mapping relationship between forces and torques presented in Equation (2), we obtain the following estimation formula for the external endeffector force:

$$F_{ext} = J(q)^{T+} (\tau_{obs} - \tau_{inv}) \tag{6}$$

where  $au_{obs}$  denotes the observed joint torques and  $au_{inv}$  represents the expected joint torques computed from the robot dynamics under the current motion state. In addition, the pseudoinverse of the transposed Jacobian  $J(q)^T$ , is computed via singular value decomposition (SVD):  $J(q)^{T+} = V\Sigma^+U^T$  if  $J(q)^T = U\Sigma V^T$ , where  $\Sigma^+$  is obtained by taking the reciprocal of each nonzero singular value in  $\Sigma$  and transposing the result. This SVD-based formulation ensures a stable least-squares approximation even when the joint space and task space dimensionalities are not perfectly aligned.

Synchronous digital twin. Moreover, to robustly compute expected joint torques and estimate end-effector forces in real-world robots affected by factors such as sensor noise, joint friction, and mechanical compliance, we employ a synchronous digital twin framework. Specifically, a high-fidelity robotic arm model is instantiated in the MuJoCo simulator, where measured joint positions, velocities, and torques from the physical robot are continuously synchronized in real time. Exploiting MuJoCo's precise dynamics simulation, the framework computes expected joint torques  $\tau_{inv}$  and infers the corresponding external wrench  $F_{ext}$  on the end-effector. This approach obviates the need for additional high-precision force or acceleration sensors, minimizes hardware dependency, and can be generalized to different robotic platforms by simply updating the simulation model.

## C. Demonstration Data Collection with Force Feedback

In imitation learning, collecting high-quality demonstration data is essential, particularly when force signals are included as part of the model input. To ensure that demonstrations capture informative variations in force, we develop two low-cost teleoperation frameworks that enhance the operator's perception of interaction forces and improve the quality of recorded trajectories. Vibro-haptic force feedback. The first framework employs a handheld controller with vibrotactile haptic feedback, as shown in the left half of Fig. 3. Estimated end-effector forces, computed from Equation (6), are mapped to vibration cues, where the vibration frequency increases discretely with force magnitude. This design conveys qualitative information about contact intensity, enabling demonstrators to intuitively adjust their actions. As a result, the collected trajectories better reflect meaningful force variations, which strengthens the robustness of the imitation learning.

VR-based visual force feedback. In the second setup, as shown in the right half of Fig. 3, a VR headset displays the estimated force vectors in real time, effectively providing an AR-like overlay on the real environment. This allows the operator to directly perceive both the direction and magnitude of contact forces. This spatially explicit visualization helps demonstrators adapt their manipulation strategies with higher precision, leading to demonstration data that more accurately encode the relationship between forces and actions.

These complementary methods can be used individually or together, allowing users to choose the approach that fits their available hardware: the vibro-haptic method is more cost-efficient, while the VR-based visual method provides additional force-direction cues, ensuring more flexible and high-quality demonstrations.

#### V. EXPERIMENTS

Experiments comprise simulation and real-robot studies. We first describe the experimental setups and data-collection protocols, then present results with a comparative analysis.

## A. Simulation Setup

**Experimental platform.** We use MuJoCo for its efficiency and accuracy in contact-rich manipulation. The robot is a 6-DOF manipulator with a fixed rectangular peg (0.95 cm side length). The environment is a flat tabletop with a 1 cm square hole. Two RGB cameras provide local (end-effector) and global views (Fig. 5a), though cameras are not rendered in illustrations for clarity.

**Task description.** The task is peg-in-hole insertion: the end-effector must align the rectangular peg with the square hole and insert it. Only translational DoFs (x, y, z) are controlled, with orientation fixed perpendicular to the table. The hole starts slightly below and in front of the peg. Force feedback is critical, as small contacts are often unobservable visually. Fig. 5 shows the task process.

**Data collection.** Unlike conventional scripted datasets, we collect demonstrations using a haptic-feedback handheld controller to capture force-guided corrective behaviors. MuJoCo visualizes contact forces directly (Fig. 6). Each trajectory begins from a fixed end-effector pose, with the hole center randomly offset by 1 mm. We collected 150 trajectories: 50 single-shot insertions without contact and 100 with corrective motions. Each trajectory includes synchronized RGB images, Cartesian positions, joint torques, end-effector wrenches, and operator commands.

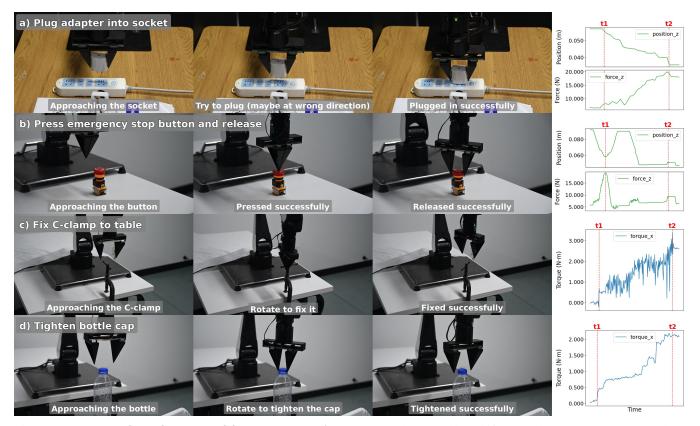


Fig. 4: **Process and force feedback of four demonstration tasks.** Peaks or sudden shifts mark key contact events, such as pressing, releasing, locking, or tightening, reflecting the different physical interactions involved.

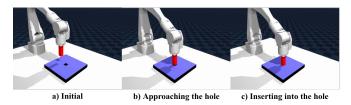


Fig. 5: Simulation scene setup and step-by-step process of the peg-in-hole task.

## B. Real World Setup

**Experimental platform.** We use a 6-DOF AIRBOT Play robotic arm as the manipulator. A charging plug is rigidly mounted on the end-effector. The arm manipulates on a white tabletop with a fixed power socket. Two RGB cameras are placed in front of and to the left of the socket to provide comprehensive scene observation. The experimental setup is shown in Fig. 4 a).

**Task description.** The task is to insert the charging plug into the socket. As in simulation, the end-effector orientation is constrained to be perpendicular to the table, and only x, y, z translation is controlled. At initialization, the end-effector is above but laterally offset from the socket by approximately 2 cm. Fig. 4 a) illustrates the task completion process.

**Data collection.** Unlike simulation, contact forces cannot be visualized directly, so the demonstrations are collected using the VR-based teleoperation interface. And the socket position remains fixed across trajectories. We collect 30

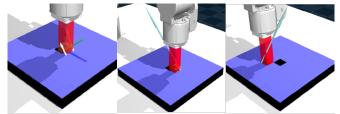


Fig. 6: **Visualization of contact forces in the simulator.** The direction and length of the blue arrow represent the direction and relative magnitude of the force at the contact point.

trajectories in total: 10 single-shot insertions without contact and 20 trajectories with corrective actions guided by force feedback. The recorded data modalities mirror those in simulation.

#### C. Experimental Results

To evaluate the relative utility of estimated end-effector force versus raw joint torque as proprioceptive contact cues for robotic insertion tasks, we conduct a controlled ablation study in both simulation and real-world settings. All model architectures and training hyperparameters (detailed in Table I) are held constant across variants, and the only difference lies in the choice of proprioceptive input: either joint torque or estimated end-effector force. In addition to this contact cue, all models observe the end-effector position and RGB images to ensure fair comparison. Each experimen-

TABLE I: Model hyperparameters in the experiment.

Hyperparameter	Value
Learning rate	2e-5
Batch size	16
# Encoder layers	4
# Decoder layers	7
Feedforward dimension	3200
Hidden dimension	512
# Heads	8
Chunk size	25
Beta	10
Dropout	0.1

TABLE II: Success rate under different proprioceptive inputs. All models take end-effector position and RGB images as inputs; they differ only in the proprioceptive cue: joint torque or estimated end-effector (EE) force.

Platform	Proprioceptive Input	Success Rate (%)
Simulation	EE position	68.0
	EE position + joint torque	80.0
	EE position + EE force	90.0
Real robot	EE position	46.7
	EE position + joint torque	63.3
	EE position + EE force	80.0

tal condition, for both simulation and real-world trials, was evaluated over 50 and 30 independent trials, respectively, to calculate the success rate, summarized in Table II.

The results reveal three critical insights:

First, the baseline model using only end-effector position and vision achieves moderate success (68% in simulation, 46.7% on the real robot), indicating that geometric and visual cues alone are insufficient for reliable insertion. This is especially evident in the real-world setting, where unmodeled dynamics and sensor noise increase failures. Insertion tasks have extremely low tolerance for positional error: even minor misalignments can cause contact-induced jamming. Without force feedback, the policy cannot distinguish between free and obstructed motion, preventing corrective actions and explaining the higher failure rate in the position-only condition.

Second, augmenting proprioception with any form of force/torque signal, whether raw joint torque or estimated end-effector force, substantially improves performance. In simulation, success increases by 12–22 percentage points (pp); on the real robot, gains are even larger (16.6–33.3 pp). This confirms that force feedback is indispensable for contact-rich manipulation, enabling the policy to perceive interaction states and modulate behavior accordingly.

Third, and most importantly, estimated end-effector force consistently outperforms raw joint torque. In simulation, success increases from 80.0% to 90.0% (+10 pp); on the real robot, from 63.3% to 80.0% (+16.7 pp). End-effector force is spatially localized and task-aligned, activating only upon contact at the tool tip and providing a clean, directional signal correlated with insertion progress. In contrast, raw joint torque is influenced by inertial, gravitational, and frictional effects along the kinematic chain, making it noisy and ambiguous, especially under dynamic motion or on real hardware. The larger performance gap on the real robot

further shows that our end-effector force estimation module enables the policy to extract more robust and task-relevant contact features.

In summary, these results demonstrate that:

- (i) Force perception is critical in insertion tasks, where low error tolerance means that, without it, policies are blind to collisions and unable to recover from jamming;
- (ii) Not all types of force signals exert the same effect: compared with joint torques, the estimated end-effector force provides superior task performance by delivering spatially precise and semantically meaningful contact feedback.

## D. End-effector Force Estimation Demonstration

To further demonstrate the effectiveness of end-wrench information, we conducted real-world experiments to test the changes in end-force/torque on a variety of contact-rich tasks, including the socket insertion and three additional tasks.

**Plug adapter into socket.** The socket is fixed on the table, and the robotic arm attempts to insert the gripping adapter. At time  $t_1$ , the Z-axis external force increases slightly, indicating that insertion has started and is proceeding correctly. The force then rises gradually, reaching a peak at time  $t_2$ , indicating that the adapter has been successfully plugged in.

**Press emergency stop button and release.** The emergency stop button is fixed on the table. The robotic arm, with a closed gripper, presses the button from above. The gripper then opens to grasp the button before the arm rotates to twist it, allowing the button to be released. At time  $t_1$ , the Z-axis external force reaches a peak, indicating the button has been pressed. At time  $t_2$ , the Z-axis force suddenly increases, indicating the button has been released and the spring exerts an opposing reaction force.

Fix C-clamp to table. The C-clamp is placed on the table, and the robotic arm grasps the lever and rotates to secure it. At time  $t_1$ , the X-axis external torque increases suddenly, indicating the arm has started applying wrist rotation. From  $t_1$  to  $t_2$ , the arm continues rotating its wrist, while the X-axis torque gradually rises and reaches a peak at  $t_2$ , indicating the C-clamp is securely fixed.

**Tighten bottle cap.** The bottle is fixed on the table, and the robotic arm grips the cap and rotates to tighten it. At time  $t_1$ , the X-axis external torque increases suddenly, indicating the arm has started applying wrist rotation. From  $t_1$  to  $t_2$ , the arm continues rotating its wrist, while the X-axis torque gradually rises and reaches a peak at  $t_2$ , indicating the cap is fully tightened.

These experiments indicate that changes in end-effector forces and torques effectively capture key phases of contactrich tasks. Peaks or sudden shifts in these values reliably indicate the completion of interactions, highlighting the importance of end-effector force feedback for real-time perception and state estimation in robotic manipulation.

# VI. CONCLUSION

We present FILIC, a dual-loop force-aware imitation learning framework that unifies a high-level Transformer policy

with a low-level impedance controller for compliant, contactrich manipulation. FILIC leverages a sensorless, Jacobian-based end-effector force estimation, fused with RGB observations, while the inner impedance loop ensures torque-level compliance. We develop two novel complementary data collection approach using VR visual force feedback and haptic vibrations, enabling efficient and safe force-aware demonstrations. Experiments on a 6-DoF robot show that replacing raw joint torque with estimated end-effector force improves success rates and enables safer, more reliable contact handling, highlighting that explicit force representation and inner-loop compliance complement modern IL policies without requiring expensive F/T sensors.

**Future work**. Limitations include reliance on accurate kinematics/dynamics, fixed end-effector orientation constraints, and a limited evaluation set. Future work will expand task diversity to full 6-DoF control, integrate tactile or event cameras, enable real-time stiffness modulation with optimal control, and explore VLA-based instruction following with force-aware prompting and multi-task datasets.

#### REFERENCES

- Z. Zhang, H. Xu, Z. Yang, C. Yue, Z. Lin, H.-a. Gao, Z. Wang, and H. Zhao, "Ta-vla: Elucidating the design space of torque-aware visionlanguage-action models," arXiv preprint arXiv:2509.07962, 2025.
- [2] X. Li, B. Guo, J. Han, and X. Zhang, "Research on hybrid force/position control method for robot peg-in-hole assembly," *Advances in Mechanical Engineering*, vol. 17, no. 5, p. 16878132241304254, 2025.
- [3] W. Hu, B. Huang, W. W. Lee, S. Yang, Y. Zheng, and Z. Li, "Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing," *Robotics and Autonomous* Systems, vol. 186, p. 104904, 2025.
- [4] S. Stepputtis, M. Bandari, S. Schaal, and H. B. Amor, "A system for imitation learning of contact-rich bimanual manipulation policies," in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 11810–11817, IEEE, 2022.
- [5] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," arXiv preprint arXiv:2304.13705, 2023.
- [6] 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," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [7] D. Wang, J. Guo, C. Sun, M. Xu, and Y. Zhang, "A flexible concept for designing multiaxis force/torque sensors using force closure theorem," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 7, pp. 1951–1959, 2013.
- [8] J. H. Kang, S. Joshi, R. Huang, and S. K. Gupta, "Robotic compliant object prying using diffusion policy guided by vision and force observations," *IEEE Robotics and Automation Letters*, 2025.
- [9] C. C. Beltran-Hernandez, D. Petit, I. G. Ramirez-Alpizar, T. Nishi, S. Kikuchi, T. Matsubara, and K. Harada, "Learning force control for contact-rich manipulation tasks with rigid position-controlled robots," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5709–5716, 2020
- [10] J. DelPreto, J. I. Lipton, L. Sanneman, A. J. Fay, C. Fourie, C. Choi, and D. Rus, "Helping robots learn: a human-robot master-apprentice model using demonstrations via virtual reality teleoperation," in 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 10226–10233, IEEE, 2020.
- [11] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.
- [12] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al., "Openvla: An open-source vision-language-action model," arXiv preprint arXiv:2406.09246, 2024.

- [13] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al., "π0: A vision-languageaction flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550," arXiv preprint ARXIV.2410.24164.
- [14] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al., "\u03c40.5: a vision-language-action model with open-world generalization," arXiv preprint arXiv:2504.16054, 2025.
- [15] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, H. Cui, et al., "Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data," arXiv preprint arXiv:2505.03233, 2025.
- [16] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, et al., "Smolvla: A vision-language-action model for affordable and efficient robotics," arXiv preprint arXiv:2506.01844, 2025.
- [17] K. Black, M. Y. Galliker, and S. Levine, "Real-time execution of action chunking flow policies," arXiv preprint arXiv:2506.07339, 2025.
- [18] D. Driess, J. T. Springenberg, B. Ichter, L. Yu, A. Li-Bell, K. Pertsch, A. Z. Ren, H. Walke, Q. Vuong, L. X. Shi, et al., "Knowledge insulating vision-language-action models: Train fast, run fast, generalize better," arXiv preprint arXiv:2505.23705, 2025.
- [19] C. Pan, K. Junge, and J. Hughes, "Vision-language-action model and diffusion policy switching enables dexterous control of an anthropomorphic hand," arXiv preprint arXiv:2410.14022, 2024.
- [20] N. Hogan, "Impedance control: An approach to manipulation: Part ii—implementation," *Journal of dynamic systems, measurement, and control*, vol. 107, no. 1, pp. 8–16, 1985.
- [21] A. M. Abdullahi, A. Haruna, and R. Chaichaowarat, "Hybrid adaptive impedance and admittance control based on the sensorless estimation of interaction joint torque for exoskeletons: a case study of an upper limb rehabilitation robot," *Journal of Sensor and Actuator Networks*, vol. 13, no. 2, p. 24, 2024.
- [22] G. Vitrani, S. Cortinovis, L. Fiorio, M. Maggiali, and R. A. Romeo, "Improving the grasping force behavior of a robotic gripper: model, simulations, and experiments," *Robotics*, vol. 12, no. 6, p. 148, 2023.
- [23] Z. Cong, A. Honglei, C. Wu, L. Lang, Q. Wei, and M. Hongxu, "Contact force estimation method of legged-robot and its application in impedance control," *IEEE Access*, vol. 8, pp. 161175–161187, 2020.
- [24] S. S. Kotha, N. Akter, S. H. Abhi, S. K. Das, M. R. Islam, M. F. Ali, M. H. Ahamed, M. M. Islam, S. K. Sarker, M. F. R. Badal, et al., "Next generation legged robot locomotion: A review on control techniques," *Heliyon*, vol. 10, no. 18, 2024.
- [25] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 572–587, 2024.
- [26] H. Nguyen and H. La, "Review of deep reinforcement learning for robot manipulation," in 2019 Third IEEE international conference on robotic computing (IRC), pp. 590–595, IEEE, 2019.
- [27] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in 2022 International Conference on Robotics and Automation (ICRA), pp. 8298–8304, IEEE, 2022.
- [28] W. Qi, H. Fan, C. Zheng, H. Su, and S. Alfayad, "Human-like dexterous grasping through reinforcement learning and multimodal perception," *Biomimetics*, vol. 10, no. 3, p. 186, 2025.
- [29] A. Yu, A. Foote, R. Mooney, and R. Martín-Martín, "Natural language can help bridge the sim2real gap," arXiv preprint arXiv:2405.10020, 2024
- [30] Y. Wu, Z. Chen, F. Wu, L. Chen, L. Zhang, Z. Bing, A. Swikir, S. Haddadin, and A. Knoll, "Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation," in 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 11831–11837, IEEE, 2025.
- [31] B. Zhou, R. Jiao, Y. Li, X. Yuan, F. Fang, and S. Li, "Admittance visuomotor policy learning for general-purpose contact-rich manipulations," *IEEE Transactions on Industrial Electronics*, 2025.
- [32] Z. Zhou, X. Yang, and X. Zhang, "Variable impedance control on contact-rich manipulation of a collaborative industrial mobile manipulator: An imitation learning approach," *Robotics and Computer-Integrated Manufacturing*, vol. 92, p. 102896, 2025.
- [33] A. B. Tahmaz, R. Prakash, and J. Kober, "Impedance primitive-augmented hierarchical reinforcement learning for sequential tasks," in 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 10973–10979, IEEE, 2025.

- [34] Z. Zhang, Y. Wang, Z. Zhang, L. Wang, H. Huang, and Q. Cao, "A residual reinforcement learning method for robotic assembly using visual and force information," *Journal of Manufacturing Systems*, vol. 72, pp. 245–262, 2024.
- [35] S. Stepputtis, M. Bandari, S. Schaal, and H. B. Amor, "A system for imitation learning of contact-rich bimanual manipulation policies," in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 11810–11817, IEEE, 2022.
- [36] M. Kim, S. Niekum, and A. D. Deshpande, "Scape: Learning stiffness control from augmented position control experiences," in *Conference on Robot Learning*, pp. 1512–1521, PMLR, 2022.
  [37] J. Luo, E. Solowjow, C. Wen, J. A. Ojea, A. M. Agogino, A. Tamar,
- [37] J. Luo, E. Solowjow, C. Wen, J. A. Ojea, A. M. Agogino, A. Tamar, and P. Abbeel, "Reinforcement learning on variable impedance controller for high-precision robotic assembly," in 2019 international conference on robotics and automation (ICRA), pp. 3080–3087, IEEE, 2019.