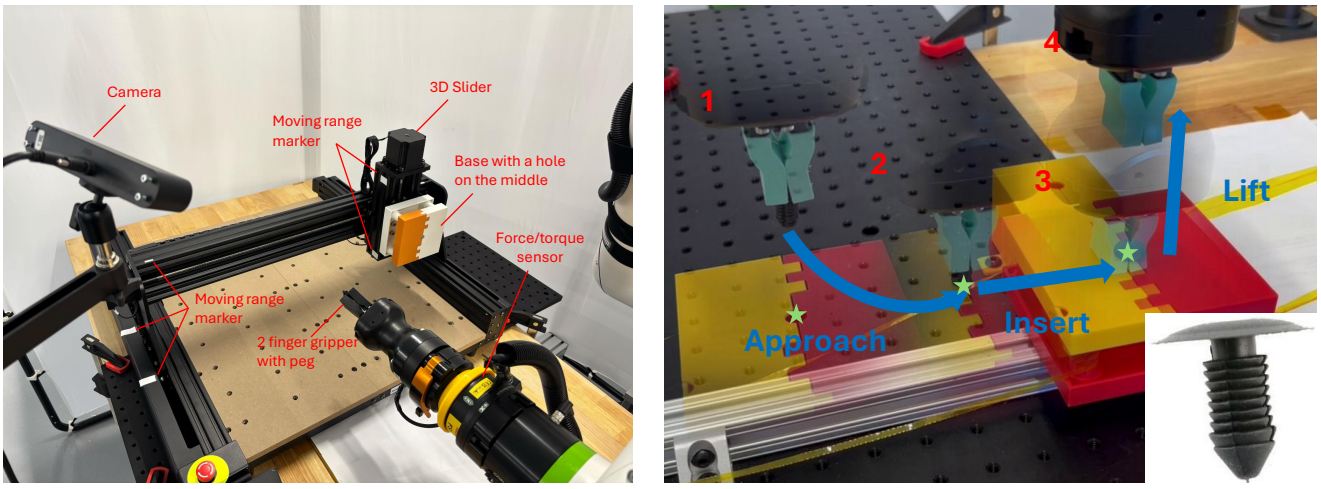
Vision-Force Admittance Learning for Peg Insertion into a Movable Hole

Yuzhong Chen[1], Yongqing Liang[1], Yunzhi Xu[2], Irving Fang[1], Chase Kidder[2], Hui-ping Wang[2], Raihan Haque[2], Yubiao Zhang[2], Chen Feng[1,✉]

https://ai4ce.github.io/VFAL/



(a) Experiment Setup (b) Insertion Process

Fig. 1: Overview of our real-world experiments. (a) shows the mobile base insertion setup. In (b), when the base moves with unknown motion, VFAL drives the robot to an approximate pre-insertion pose, then performs precise insertion using asynchronous force and visual feedback. After insertion, the robot lifts the end-effector.

***Abstract*—Precise manipulation in dynamic environments, whether induced by a mobile robot base or a target with unknown motion, remains a major challenge in robotics. Manipulation in dynamic environments introduces substantial uncertainty, which fundamentally conflicts with the tight precision requirement of precise tasks such as peg-in-the-hole. We propose a Vision-Force Admittance Learning (VFAL) framework that fuses asynchronous visual feedback with a high-frequency force-based model, using visual pose estimations as a regularization term. VFAL adapts insertion strategies online to dynamic motion while maintaining millimeter-level precision. To obtain robust, low-frequency pose information, we employ state-of-the-art vision foundation models for visual pose estimation. Additionally, we incorporate failure recovery mechanisms to enhance overall robustness. We validate our approach in real-world experiments, demonstrating high success rates and strong adaptability to various pegs and dynamic environments.**

I. Introduction

Precise manipulation is a long-standing challenge in robotics and has been extensively studied in static settings. Most existing approaches assume fixed-base robots interacting with stationary targets, where uncertainty is primarily from unknown target pose and contact behavior, and high accuracy is attainable. In contrast, precise manipulation in dynamic environments, such as mobile manipulation or part installation on moving production lines, has received comparatively limited attention. The key difficulty arises from additional uncertainty induced by relative motion between the robot and the target, which greatly complicates accurate execution. In this work, we study precise manipulation in dynamic environments using peg-in-movable-hole as a representative task. This scenario amplifies the challenge by coupling motion with high accuracy requirements and the need for real-time adaptation.

Achieving success in such settings requires robot systems that combine high precision with strong adaptability to environmental changes. Prior work on precise manipulation has demonstrated high precision using impedance control [1, 2], compliant strategies [3, 4], and learning-based force regulation [5, 6, 7], enabling tasks such as peg-in-hole insertion and assembly [8, 9]. However, these methods largely remain limited to static or quasi-static scenarios [3, 10]. Conversely, dynamic manipulation techniques emphasize robustness to motion through visual servoing [11], prediction [12, 13], or reactive control [14], yet typically offer limited precision [14, 15].

Humans, by contrast, achieve precise manipulation in dynamic environments by flexibly integrating multiple sensing modalities: combining low-frequency, high-latency visual perception with high-frequency, low-latency force feedback to

Manuscript received: June 6, 2026; Revised: August 14, 2026; Accepted: August 24, 2026.

This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers comments.

[1]New York University, Brooklyn, NY 11201, USA

[2]General Motors, Warren, MI 48073, USA

✉ Corresponding author (cfeng@nyu.edu). The authors gratefully acknowledge the collaboration, support, resources, and technical expertise provided by the General Motors Autonomous Robotics Center (ARC) and the Robotics Intelligence team, which were instrumental to this work.

Digital Object Identifier (DOI): see top of this page.

achieve both adaptability and precision [16, 17]. This motivates the development of robotic methods capable of fusing asynchronous modalities in a complementary manner.

A further practical challenge lies in hardware accessibility. Many robot arms lack direct force control and instead rely on position control. Implementing force control indirectly on such systems often introduces latency and instability [2]. Moreover, while pose is exploitable enough by vision-based models, force information is harder to incorporate into visual methods. These factors highlight the need for methods that can effectively utilize force feedback on position-controlled robots while still benefiting from visual priors.

To address these challenges, we present VFAL, a multimodal manipulation framework that asynchronously fuses vision and force for dynamic, precise insertion. VFAL builds upon Linear Model Learning (LML) [7] for efficient online adaptation during insertion. Specifically, VFAL comprises two key components: (i) a vision-based pose estimation module that provides robust, CAD-based pose estimates at low frequency, and (ii) a force-based model with asynchronous visual referencing, in which model parameters are updated online using control and force feedback starting from a bootstrap initialization. Force-based predictions are regularized using the asynchronously updated pose. To improve robustness, VFAL incorporates recovery strategies for common failure modes.

We claim these contributions:

1) We introduce a novel multimodal insertion framework that asynchronously fuses force and vision to enhance precision and adaptability in peg-in-movable-hole tasks.
2) We propose an asynchronous modality fusion pipeline that leverages predicted target pose as a regularization term for high-frequency admittance-based control.
3) We deploy VFAL on a position controlled robot and demonstrate robust peg insertion into a moving base.

II. Related Works

A. Peg-in-Hole Methods

The peg-in-hole problem has been extensively studied and remains popular in the robotics community due to its high precision requirements and widespread use in real-world applications. Early work primarily relied on model-based methods, including hybrid force/position control [3, 1] and analytical contact modeling [8, 10]. To address large initial misalignment, search-based [4, 10] strategies have been employed to locate holes initially outside the effective force-sensing region. More recently, learning-based approaches have gained attention for their strong generalization capability [18, 19] and for incorporating multi-modal sensing [18, 20]. Beyond rigid peg-in-hole settings, some studies have also explored assembly with deformable objects [21, 22].

However, previous methods mainly assume a static environment [4, 8, 20], where the hole is fixed but its position may be initially unknown. In contrast, our research focuses on a dynamic environment in which the hole position may change over time. This setting requires a model capable of responding to environmental changes while maintaining global sensing capability to handle large initial misalignment.

B. Force-Based Robot Manipulation

Force-based methods are well suited for precise manipulation due to their ability to operate at high frequency with low latency [19], and their higher sensor precision in contact-rich manipulation compared to visual sensors. Tactile sensing provides rich contact information [23, 24], but tactile sensors face durability challenges in industrial contexts, where long-term usage, chemical exposure, and temperature variation are common [25]. Additionally, leveraging tactile data typically requires complex neural networks, whose computation latency limits their capability in high-frequency scenarios [26].

In contrast, force-torque sensors yield low-dimensional, high-frequency signals that are lightweight to process. Classical control methods such as Model Predictive Control (MPC) have been widely applied in this setting [27, 28]. However, MPC requires modeling of dynamics and lacks generalizability across varying tasks [29]. Learning-based force control offers an alternative: these methods generalize without explicit dynamics models [19, 9, 6]. Yet large models often fail to fully exploit the fast feedback provided by force sensors, while smaller models struggle in complex scenarios—such as mobile peg-in-hole insertion with deformable fasteners—where variability challenges robustness.

Online learning presents a promising path for adapting to dynamic and uncertain conditions [30, 7, 31]. Nevertheless, adaptation during the early stages of contact is often unstable, leading to frequent failures [30, 7], or requiring ground-truth demonstrations to guide the model [31], which are difficult to obtain for precise manipulation in dynamic environments. These limitations motivate the use of global, but potentially stale and imprecise information to assist local force-based adaptation in challenging mobile assembly settings.

C. Vision-Force Fusion for Robotic Manipulation

Vision provides rich global context and improves robustness to environmental variation [12, 32, 33]. However, the high latency of visual processing hinders its effectiveness in dynamic and precise tasks where millisecond-level reactivity is preferred [32, 34].

Hybrid approaches have attempted to combine vision and force sensing to balance global context with precision [35, 36, 37]. While conceptually appealing, these methods often encounter practical mismatches: force sensors operate at high frequency with low dimensionality, whereas vision are slower and high-dimensional. As a result, fusion pipelines may sacrifice responsiveness [36, 25] or underutilize vision during critical contact stages [35]. Although [38] introduced an asynchronous module to mitigate these issues, the method has not been validated with high-frequency sensing. Moreover, existing data collection methods are too slow to be precise and often fail to capture robot dynamics in the collected data, which are critical for achieving high precision in dynamic environments. This gap highlights the need for model-based asynchronous fusion strategies that retain the advantages of both vision and force, ensuring neither modality is compromised.

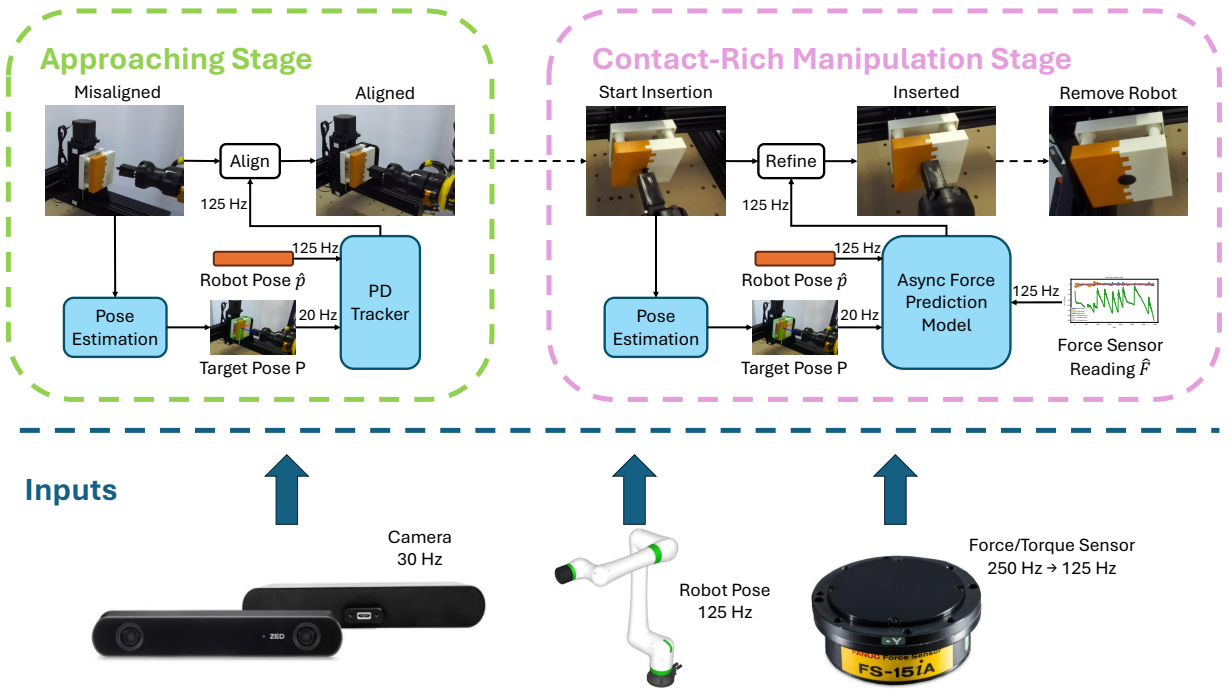


Fig. 2: **Algorithm Workflow.** A two-stage pipeline is designed for the manipulation process. In the first stage, before making contact with the base, the robot approaches the target using the pose estimation result. In the second stage, once contact is detected via the force sensor, the model predicts the optimal next step control based on the current state and force readings, while referencing the asynchronously updated pose estimation as a regularization term.

III. Method

A. Problem Overview

The peg insertion process can be divided into two stages, depending on whether contact occurs. The first stage is the approaching stage, in which the robot arm is guided to the insertion point using a vision-based method that provides global information. The second stage is the contact-rich manipulation stage, where the robot uses force-based control to precisely complete the installation at high frequency while asynchronously utilize visual information. The main challenge in this task is adapting to large target velocity and behaving robustly under target motion shifts while maintaining high precision. We address the challenge by a two stage multi-modality insertion pipeline in which vision dominant approaching is given a larger precision tolerance and residual errors are refined via asynchronous vision–force fusion during the contact-rich insertion stage.

B. Multi-modality Insertion Pipeline

As shown in Fig. 2, our model follows a two-stage scheme for peg insertion in a dynamic environment:

The approaching stage moves the robot from a predefined position to roughly contact the moving hole at an unknown position. To balance precision and stability, most existing robot arms employ a PD controller. These controllers perform well when moving toward a distant target given sufficient time. However, when tracking a target with unknown motion at high precision, the robot must react to small per-cycle changes at high frequency. In such cases, existing robot controllers will have a slow response compared to the robot control cycle time. In our experiments, this latency can cause a 1–2 cm mismatch between the robot and a moving target.

To address this issue, we design a cross-coupled PD tracker on top of a position control based robot controller which can accept target velocity as input. Given the current robot pose $\hat{p}_{i-1}$ and the pose estimation result P_i from our vision based pose estimation, adapted from FoundationPose [39] with modifications for dynamic environments. The tracker outputs the expected robot pose $p_i^{expected}$ and velocity $v_i^{expected}$ for robot controller. We update P_i using a Kalman filter to predict the target pose p_i and velocity v_i at the execution time of the current control cycle. We compute $p_i^{expected}$ and $v_i^{expected}$ as:

$$\hat{p}_i^{expected} = p_i + K_{vp} v_i + K_{pp}(p_i - \hat{p}_{i-1})$$

$$v_i^{expected} = v_i + K_{pv}(p_i - \hat{p}_{i-1})$$

This enables the robot to track the hole without large error once a rough alignment in the insertion direction established. We monitor the force reading at every control cycle.

The contact-rich manipulation stage begins once the force sensor detects a contact force exceeding a threshold during the approaching stage. We employ a force-based model with asynchronous visual referencing to reduce residual alignment errors from the approaching stage and to adapt to target motion during insertion. The objective is to maintain contact forces primarily along the insertion direction while minimizing the positional discrepancy between the estimated target pose and the executed robot pose after the current control cycle.

We denote the control u_i for step i as the offset position from current pose $\hat{p}_{i-1}$ to the next pose $\hat{p}_i$. We normalize this control to maintain a desired insertion speed along the

z-axis, which yields a more stable and consistent motion in the intended insertion direction.

C. Force-based Model with Async Visual Referencing

This module primarily operates during the contact-rich manipulation stage, where the estimated base pose is used as a reference input to a force-based decision-making process. We assume insertion proceeds downward along the z-axis, and we define the force along the insertion direction as positive.

Force-Based Adaptive Contact Modeling. We define the system state s_i as a 19-dimensional vector that concatenates the robot pose, force sensor readings, control input, and a bias term: $\left[\hat{p}_{i-1},\, \hat{F}_{i-1},\, u_i,\, 1\right]^\top$. We use a linear model G, which is widely used for position-force transformation in static environments, to transform the current system state s_i into the predicted force at the next timestep, denoted as $F_i^{\text{predicted}}$. We initialize the force prediction model as a 6×19 matrix composed of a 6×6 zero matrix for pose, two 6×6 identity matrices for force sensor reading and control, and a 6×1 zero vector for bias. This initialization reflects the intuition of bootstrapping with an admittance-like prior. During manipulation, the robot is governed by a closed-loop controller.

To achieve successful insertion, the model G minimizes contact forces in all directions orthogonal to the z-axis. Along the z-axis, positive in the insertion direction, as the contact force during insertion is typically opposite to the motion and therefore negative, we move the robot toward target by subtracting a small value $\delta = 1$ from the current force reading $\hat{F}^z_{i-1}$ until a predefined threshold $F_{\max}$, a negative value because opposite to the insertion direction, is reached. The expected force after a control cycle is

$$F_i^{\text{expected}} = \begin{cases} \mathbf{0}, & \text{orthogonal directions to } z, \\ \max\left(\hat{F}^z_{i-1} - \delta,\, F_{\max}\right), & \text{insertion direction } z. \end{cases} \tag{1}$$

In dynamic environments that often exhibit nonlinear behavior and violate quasi-static assumptions, the linear force prediction model G may no longer be valid. To address this issue, we reformulate nonlinear force behavior into a linearized quasi-static model by introducing state-of-the-art high-frequency online learning for linear models to perform the insertion task [7]. Specifically, the contact model is updated from G_{i-1} to G_i given the current state s_i and the next-step force sensing $\hat{F}_i$ using a Kalman filter. We discretize the contact-rich manipulation process into small timesteps with a limited motion range and update the transformation matrix at each step as $G_{i+1} = U(G_i, \hat{F}_i, s_i)$, where U is the update function defined in [7] that updates G based on the robot state s_i and current force/torque sensor readings. This yields a stepwise linear relationship within each small time or action interval: $F_{i+1} = G_{i+1}\, s_{i+1}$. Suppose the underlying environment dynamics evolve only slightly within a single control cycle, the force-prediction model obtained at timestep i remains close to that at timestep $i+1$. Although each timestep employs a linear model, sequential updates across control cycles allow a globally nonlinear force behavior over longer horizons.

Asynchronous Visual Referencing. In real-world settings, limited control frequency and slow learning speed, caused by discrepancies between model assumptions and environment dynamics, can prevent online learning from adapting to a moving target or target motion shift quickly enough to eliminate undesired horizontal forces. These forces may arise from lag between the robot's current pose and the target pose, or from overshoot in the previous motion direction when motion shifts. Such horizontal forces can further induce large lateral motions, which may violate the assumptions required for linear force modeling. In addition, force sensing is only reliable within a limited contact range, leaving this range may make force measurements unreliable. To address these issues, we introduce a visually estimated pose P as a regularizer, transforming the problem from adapting to target motion into making minor adjustments to pose estimation error. This helps maintain the local linearity assumption and restores reliable force sensing.

Due to high computational latency and the low frequency of sensing and inference, a significant temporal gap exists between image capture time and force-based control reference. When a new pose estimate P_j with capture time $t_0(j)$ becomes available to the force model, we update a Kalman filter that assumes the target object moves with an unknown constant velocity. During control prediction, the Kalman filter estimates the target pose p_i and velocity v_i, and the terminal velocity of the control cycle is set to v_i.

Using the imperfect visual estimation p_i as a prior, we penalize the positional discrepancy between the expected pose after executing the control, $\hat{p}_i^{\text{expected}}$, and p_i via $\|\hat{p}_i^{\text{expected}} - p_i\|_2^2$. We then formulate the control objective as a weighted sum of force-tracking and pose-regularization terms with weight λ:

$$\arg\min_{u_i} \left\|F_i^{\text{expected}} - F_i\right\|_2^2 + \lambda \left\|\hat{p}_i^{\text{expected}} - p_i\right\|_2^2. \tag{2}$$

The weight λ controls the trade-off between force-based exploration and adherence to the visual prior. When the target exhibits fast but predictable motion, a larger λ constrains exploration and maintains alignment with the estimated pose. Conversely, when the target is quasi-static or its motion is highly unpredictable, a smaller λ places greater emphasis on contact feedback.

In practice, we found that a fixed value of $\lambda = 0.1$ provides robust performance across a wide range of target motions. Specifically, this setting works well for targets exhibiting approximately linear motion with small variance, as well as trajectories with occasional direction changes along any axis. Despite variations in target velocity and motion patterns, $\lambda = 0.1$ consistently balances adherence to the visual prior and force-based correction, and therefore used throughout our experiments unless otherwise noted.

We mark insertion as complete when the current force reading along the z-axis exceeds the threshold $F_{\max}$; the robot then lifts. A diagram of VFAL during the contact-rich manipulation stage is shown in Fig. 3.

We also introduce two simple but effective failure recovery methods to enhance robustness in dynamic environments. The first addresses loss of tracking. If the camera fails to track the object, the robot retreats to a safe position until tracking is restored by the pose estimation module. The second method

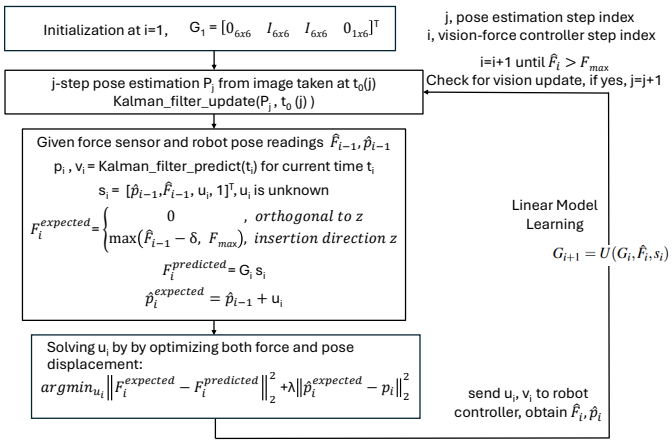


Fig. 3: **Asynchronous Vision–Force Controller.** In each control cycle, the force-based model updates itself using force/torque feedback, the robot's previous control inputs, and its previous state. When available, visual information is incorporated asynchronously. The model predicts the next-step control by minimizing the discrepancy between the expected and the sensed force and pose, using the current robot pose, force/torque readings, and visual input.

addresses inaccurate initial pose during contact-rich insertion. Erroneous or outdated pose estimates, or cases where the robot fails to accurately track the target, may cause the insertion to fail during the approaching stage. These issues are detected by monitoring force peaks that indicates successful fastener engagement. If the force exceeds a threshold $\hat{F}_{\max}$ without a corresponding force peak, the attempt is classified as a failure. In such cases, the robot moves upward and reinitializes the installation process.

IV. Experiment

We evaluate methods in dynamic environments where pegs are inserted into moving bases with holes. We further evaluate additional experimental settings quantitatively. More settings are explored through the video demonstrations.

A. Experiment Setup

We use pushnuts, a class of deformable fasteners, as experimental objects. Specifically, we use Christmas-tree pushnuts (see video demo) with conical, stepped geometries common in industrial and consumer settings. These pushnuts are inserted into different bases. All experiments use position control with a capped z-axis speed during contact-rich insertion.

We evaluate performance using two metrics. The success rate measures the possibility of successful insertions. A trial is successful if insertion completes with at most one fin outside the hole when the z-axis force reaches $F_{\max}$. We exclude failures occurring during the shared approaching stage before hole contact. These failures are considered external to contact-rich insertion. We adapt the full 6D wrench MSE metric from [40] to obtain an aggregate view of adaptation quality, and tailor it to our dynamic insertion task as $\frac{1}{trials \cdot steps} \sum_{trials} \sum_{steps} \|\mathbf{m} \odot \mathbf{F}_{observed}\|^2$, $\mathbf{m} = [1,1,0,1,1,1]$. Since the force and torque components are summed in a single squared norm, this quantity serves only as an aggregate indicator of adaptation quality and does not carry a direct physical interpretation. An ablation evaluating the force and torque components separately is provided in Sec. IV-C4. Only successful trials are included, with 10 trials per method.

As shown in Fig. 4, we design a TPU-based gripper with an internal undercut bowl. Horizontal force causes visible deformation and eventual ejection, providing clear force and failure indication while protecting the system.

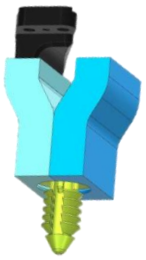


Fig. 4: **TPU-based gripper**. It serves as an indicator of force and manipulation status visualizer through its deformation.

The experimental setup is shown in Fig. 1(a) and Fig. 5. A table-mounted RGB-D camera is calibrated to the robot using an AprilTag. AprilTags are used only for camera–robot calibration. We experiment three bases: fixed, 1D mobile, and 3D mobile. The 3D mobile base is mounted on a preprogrammed motorized 3D slider. A cross-fin pushnut is used for this setting. The 1D base moves along a linear track via deformable cables. Only robotic actuation is used for controlled evaluation. Track friction and cable elasticity introduce velocity variation under contact. A cylindrical pushnut is used for 1D experiments. All experiments use a FANUC CRX-10iA/L with an FS-15iA force/torque sensor. Video demo presents and evaluates more different kind of pegs used in the experiment. Unless noted for synchronous ablations, control and force sensing run at 125 Hz, force data is averaged from a 250 Hz sensor, and vision runs at 20 Hz.

Fixed-base insertions are used to characterize pushnut force patterns. Eight fins in cylinder Christmas tree nut (see Fig. 1(b)) generate eight force/torque peaks. These signatures enable automatic failure detection and recovery.

B. Quantitative Results and Discussions

We adapt an incremental design for both mobile bases in our experiment.

(i) **Vision**: Vision-only insertion using estimated base pose.
(ii) **Admittance**: Force-only admittance control under dynamic environment.
(iii) **Online Learning**: Admittance augmented with online learning from [7].
(iv) **Synchronous VFAL**: Vision–force fusion with synchronized sensing via blocking.
(v) **VFAL**: Proposed asynchronous vision-assisted force model learning.

Discussion on 1D and 3D environment. The 1D mobile base is cable-driven and mechanically compliant, allowing it to deform under external forces and thereby tolerate larger positional errors during insertion. In contrast, the 3D mobile base is motor-driven and significantly stiffer, such that even small pose deviations can induce large contact forces. Additionally, the cross-fin Christmas-tree pushnut used in the 3D setting has smaller fins and is therefore easier to disengage

TABLE I: Success rate and Mean Squared Error (MSE) on full wrench across different models for both fixed-base and mobile-base insertions. We mark the best performed results for each experiment in **bold**.

Description	Method	Modality	Success Rate	Full Wrench
1D Mobile Base (Cylinder Christmas tree nut)	Vision	Vision	40%	57.11($\pm$59.65)
	Admittance	Force	60%	70.96($\pm$29.78)
	Online Learning	Force	70%	80.95($\pm$42.56)
	Synchronous VFAL	Force, Vision	90%	28.50($\pm$14.11)
	VFAL	Force, Vision	**100%**	**13.95**($\pm$**5.05**)
3D Mobile Base (Cross-fin Christmas tree nut)	Vision	Vision	20%	114.60($\pm$57.30)
	Admittance	Force	10%	106.09(N/A)
	Online Learning	Force	0%	N/A(N/A)
	Synchronous VFAL	Force, Vision	60%	37.82($\pm$15.83)
	VFAL	Force, Vision	**80%**	**9.22**($\pm$**5.23**)

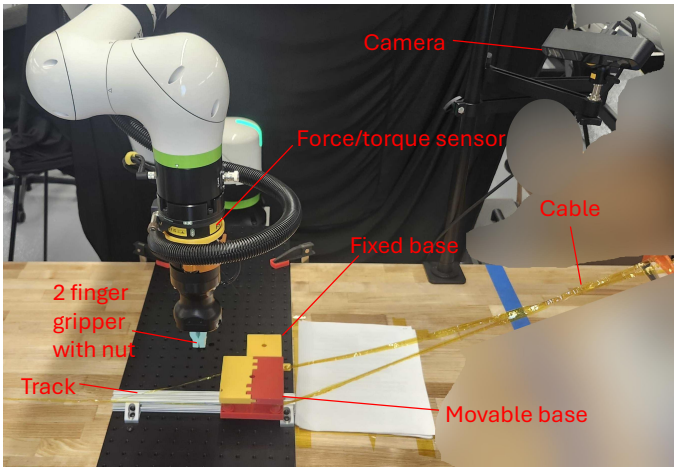


Fig. 5: **Real world experiment setup**. In our real world experiment setup, an RGB-D camera is used for pose estimation. The robot, equipped with a force/torque sensor, holds the deformable pushnut and performs the insertion task. The setup includes two types of bases: a fixed base mounted on a motherboard, and a cable-driven movable base mounted on a fixed linear track. This configuration allows us to evaluate insertion performance under both static and dynamic conditions.

than the cylindrical pushnut used in the 1D setting. As reflected in Table I, these factors make 3D insertion substantially more demanding, requiring higher manipulation precision and resulting in degraded performance across most metrics.

Discussion on vision-only model and force-only models. Vision-only methods are generally more transferable, as they leverage global context and scene-level information. However, they suffer from higher latency due to sensing and computation overhead, which can cause large errors in dynamic environments. Moreover, visual sensing is typically less precise than force/torque sensing in contact-rich manipulation. On the 1D mobile base, the system has greater tolerance to detect contact early and adjust the insertion using force feedback. As a result, both force-only methods (**Admittance** and **Online Learning**) outperform **Vision**, as shown in Table I. In contrast, on the 3D mobile base, the robot has limited tolerance to correct its pose through contact alone. In this setting, vision-only method outperform force-only methods due to their ability to provide more robust global information.

Discussion on vision-force fusion. Results on the mobile base experiments show that combining vision and force (**Synchronous VFAL** and **VFAL**) enables higher performance than either modality alone (**Vision**, **Admittance**, and **Online Learning**). Vision–force fusion benefits from precise local force feedback while retaining global visual context, leading to improved robustness and accuracy in dynamic insertion tasks.

Discussion on asynchronous. The comparison between **Synchronous VFAL** and **VFAL** further demonstrates that incorporating visual information asynchronously improves performance. By asynchronous fusion of vision into force, the controller achieves lower latency from higher control frequency, which directly translates into better task performance.

C. Ablation Studies

1) Ablation on Online Learning: To evaluate the effect of online learning, we conduct experiments on the 3D mobile base and compare **VFAL** with **VFAL without Online Learning**, as shown in Table II.

TABLE II: Ablation on Online Learning (3D Mobile Base)

Method	Success Rate	Full Wrench MSE
VFAL without Online Learning	70%	12.11($\pm$8.04)
VFAL	**80%**	**9.22**($\pm$**5.23**)

The results indicate that online learning improves both the success rate and full wrench MSE. Online learning is still preferred even if asynchronous fusion of vision and force is provided.

2) Ablation on λ Selection: As described in the method section, we set the visual prior regularization weight to $\lambda = 0.1$ in all main experiments. To validate this choice, we evaluate different values of λ on the 3D mobile base. Each setting is tested over 10 trials.

TABLE III: Ablation on λ Selection (3D Mobile Base)

Method	Success Rate	Full Wrench MSE
$\lambda = 1$	70%	73.68($\pm$50.89)
$\lambda = 0.1$	**80%**	**9.22**($\pm$**5.23**)
$\lambda = 0.01$	40%	67.89($\pm$68.73)

Among the tested values, $\lambda = 0.1$ achieves the highest success rate and the lowest force/torque error. Increasing λ to 1 degrades performance, as an overly strong visual prior suppresses reliable force feedback, which is inherently more

precise during contact. When λ is reduced to 0.01, visual guidance becomes too weak, resulting in force-dominated behavior and substantially lower performance. Nevertheless, even at $\lambda = 0.01$, the method still outperforms pure force-only baselines on the 3D mobile base.

3) Ablation on the robot controller: We conduct an ablation study on a high frequency robot controller while keeping the force sensing frequency matching the pose tracking frequency as in **Synchronous VFAL**. This experiment evaluates whether the asynchronous vision force fusion design is the main source of VFAL's performance gain over **Synchronous VFAL**.

TABLE IV: Ablation on the robot controller (3D Mobile Base)

Method	Success Rate	Full Wrench MSE
Synchronous VFAL	60%	37.82(±15.83)
High Frequency Controller	60%	54.95(±41.11)
VFAL	**80%**	**9.22**(±**5.23**)

The performance of **High Frequency Controller** matches **Synchronous VFAL**, while **VFAL** reaches a much higher performance. Raising the controller rate is not the main source of VFAL's performance gain.

4) Ablation on force/torque: We report force and torque MSE separately, alongside the full wrench MSE, to characterize contact behavior in more detail.

TABLE V: Ablation on force/torque (3D Mobile Base)

Method	Force	Torque	Full Wrench
Vision	100.32(±51.39)	14.28(±5.91)	114.60(±57.30)
Admittance	95.02(N/A)	11.07(N/A)	106.09(N/A)
Online Learning	N/A(N/A)	N/A(N/A)	N/A(N/A)
Sync VFAL	33.26(±13.99)	4.56(±1.94)	37.82(±15.83)
VFAL	**8.07**(±**4.56**)	**1.15**(±**0.68**)	**9.22**(±**5.23**)

VFAL still performs the best among all evaluated methods under both the force and torque MSE metrics.

5) Ablation on full system: We additionally report the approaching stage and the end-to-end success rate alongside the insertion results.

TABLE VI: Ablation on full system (3D Mobile Base)

Method	Approaching	Insertion	Overall
Vision	10/17	2/10	2/17
Admittance	**10/10**	1/10	1/10
Online Learning	10/12	0/10	0/12
Synchronous VFAL	10/13	6/10	6/13
VFAL	10/12	**8/10**	**8/12**

The total number of trials for each method varies, since the main contribution of this work concerns the insertion stage. Thus, we keep collecting data until 10 trials successfully enter the insertion stage. The same results are also visualized as a Sankey diagram in the real world robot demo, experiment 1.

D. Real World Robot Demo

We deploy our method on a FANUC CRX-10iA/L robot arm to qualitatively demonstrate insertion behaviors in both static and dynamic scenarios, including different pegs, different version of target motion, and failure recovery. Demonstration videos are provided in the supplementary material.

(i) **Experiment 1:** Peg insertion on a 3D mobile base under the same setup as the 3D mobile base experiments. Results for all trials are visualized at the beginning.

(ii) **Experiment 2:** Insertion of different peg types, including a pushnut and a rigid cylinder peg, into a 3D mobile base.

(iii) **Experiment 3** & **4:** Peg insertion on a 1D mobile base under 1D mobile base experiment speed (Experiment 3) and doubled speed (Experiment 4).

(iv) **Experiment 5:** Recovery from unintended panel collisions caused by tracking failures in the approach stage.

(v) **Experiment 6:** Recovery from occlusion or losing tracking, preventing potential collisions.

(vi) **Experiment 7:** Peg insertion into a fixed base mounted on a table to present ideal force behavior for insertion.

(vii) **Experiment 8:** Peg insertion into a base with unknown and noisy human-driven motion in a 2D plane.

V. Conclusion and Limitation

In this work, we develop a modality fusion framework that asynchronously combines vision and force sensor readings without reducing sensor frequency by introducing a regularization term between the estimated pose and the expected pose after motion, and by setting a target speed after each control cycle. This framework enables the robot to perform precise peg insertion into a movable hole at a high control frequency. We evaluate our framework in the real world. Our model achieves high success rate, low force error, and demonstrates strong adaptability to dynamic environments.

The proposed framework has limitation. It assumes that the peg has already been grasped. In real-world scenarios, however, the robot must first perform reliable grasping.

References

[1] M. H. Raibert and J. J. Craig, "Hybrid position/force control of manipulators," *Journal of dynamic systems, measurement, and control*, vol. 103, no. 2, pp. 126–133, 1981.

[2] O. Khatib and J. Burdick, "Motion and force control of robot manipulators," in *Proceedings. 1986 IEEE international conference on robotics and automation*, vol. 3. IEEE, 1986, pp. 1381–1386.

[3] D. E. Whitney, "Quasi-static assembly of compliantly supported rigid parts," *Journal of Dynamic Systems, Measurement, and Control*, vol. 104, no. 1, pp. 65–77, 1982.

[4] H. Park, J. Park, D.-H. Lee, J.-H. Park, and J.-H. Bae, "Compliant peg-in-hole assembly using partial spiral force trajectory with tilted peg posture," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4447–4454, 2020.

[5] 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.

[6] 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)*. IEEE, 2019, pp. 3080–3087.

[7] K. Tracy, Z. Manchester, A. Jain, K. Go, S. Schaal, T. Erez, and Y. Tassa, "Efficient online learning of contact force models for connector insertion," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 10 944–10 950.

[8] Y. Chen, K. Kimble, H. H. Qian, P. Chanrungmaneekul, R. Seney, and K. Hang, "Robust peg-in-hole assembly under

uncertainties via compliant and interactive contact-rich manipulation," *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[9] A. Negi, O. M. Manyar, D. K. V. Penmetsa, and S. K. Gupta, "Learning the contact manifold for accurate pose estimation during peg-in-hole insertion of complex geometries," *arXiv preprint arXiv:2505.19215*, 2025.

[10] S. R. Chhatpar and M. S. Branicky, "Search strategies for peg-in-hole assemblies with position uncertainty," in *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, vol. 3. IEEE, 2001, pp. 1465–1470.

[11] M. Liu, Z. Chen, X. Cheng, Y. Ji, R.-Z. Qiu, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," in *Conference on Robot Learning*. PMLR, 2025, pp. 234–257.

[12] S. Uppal, A. Agarwal, H. Xiong, K. Shaw, and D. Pathak, "Spin: Simultaneous perception interaction and navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 133–18 142.

[13] H. Xiong, R. Mendonca, K. Shaw, and D. Pathak, "Adaptive mobile manipulation for articulated objects in the open world," *arXiv preprint arXiv:2401.14403*, 2024.

[14] B. Burgess-Limerick, "An architecture for reactive mobile manipulation on-the-move," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[15] R. Yang, Y. Kim, R. Hendrix, A. Kembhavi, X. Wang, and K. Ehsani, "Harmonic mobile manipulation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 3658–3665.

[16] F. Hutmacher, "Why is there so much more research on vision than on any other sensory modality?" *Frontiers in psychology*, vol. 10, p. 481030, 2019.

[17] H. B. Helbig and M. O. Ernst, "Optimal integration of shape information from vision and touch," *Experimental brain research*, vol. 179, no. 4, pp. 595–606, 2007.

[18] W. Chen, C. Zeng, F. Sun, J. Zhang *et al.*, "Learning multiple peg-in-hole assembly skills using multimodal representations and impedance control," *Workshop on Transferability in Robotics (TRW), IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[19] T. Inoue, G. De Magistris, A. Munawar, T. Yokoya, and R. Tachibana, "Deep reinforcement learning for high precision assembly tasks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 819–825.

[20] Y. Tao, S. Chen, H. Liu, H. Gao, Y. Tao, Y. Chen, and H. Wei, "Peg-in-hole assembly method based on visual reinforcement learning and tactile pose estimation," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 1410–1417.

[21] K. Wu, R. Chen, Q. Chen, and W. Li, "Robotic assembly of deformable linear objects via curriculum reinforcement learning," *IEEE Robotics and Automation Letters*, 2025.

[22] J. Luo, E. Solowjow, C. Wen, J. A. Ojea, and A. M. Agogino, "Deep reinforcement learning for robotic assembly of mixed deformable and rigid objects," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2062–2069.

[23] T. Matsubara and K. Shibata, "Active tactile exploration with uncertainty and travel cost for fast shape estimation of unknown objects," *Robotics and Autonomous Systems*, vol. 91, pp. 314–326, 2017.

[24] Z. Zhao, W. Li, Y. Li, T. Liu, B. Li, M. Wang, K. Du, H. Liu, Y. Zhu, Q. Wang *et al.*, "Embedding high-resolution touch across robotic hands enables adaptive human-like grasping," *Nature Machine Intelligence*, pp. 1–12, 2025.

[25] J. Zhao, N. Kuppuswamy, S. Feng, B. Burchfiel, and E. Adelson, "Polytouch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion policies," 2025.

[26] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[27] S. Kleff, E. Dantec, G. Saurel, N. Mansard, and L. Righetti, "Introducing force feedback in model predictive control," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 13 379–13 385.

[28] S. Le Cleac'h, T. A. Howell, S. Yang, C.-Y. Lee, J. Zhang, A. Bishop, M. Schwager, and Z. Manchester, "Fast contact-implicit model predictive control," *IEEE Transactions on Robotics*, vol. 40, pp. 1617–1629, 2024.

[29] S. Katayama, M. Murooka, and Y. Tazaki, "Model predictive control of legged and humanoid robots: models and algorithms," *Advanced Robotics*, vol. 37, no. 5, pp. 298–315, 2023.

[30] X. Zhang, C. Wang, L. Sun, Z. Wu, X. Zhu, and M. Tomizuka, "Efficient sim-to-real transfer of contact-rich manipulation skills with online admittance residual learning," in *Conference on Robot Learning*. PMLR, 2023, pp. 1621–1639.

[31] A. T. Le, M. Guo, N. van Duijkeren, L. Rozo, R. Krug, A. G. Kupcsik, and M. Bürger, "Learning forceful manipulation skills from multi-modal human demonstrations," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7770–7777.

[32] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024.

[33] J. Luo, C. Xu, J. Wu, and S. Levine, "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning," *Science Robotics*, vol. 10, no. 105, p. eads5033, 2025.

[34] B. Li, Z. Huang, J. Ye, Y. Li, S. Scherer, H. Zhao, and C. Fu, "Pvt++: a simple end-to-end latency-aware visual tracking framework," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10 006–10 016.

[35] Z. Zhao, S. Haldar, J. Cui, L. Pinto, and R. Bhirangi, "Touch begins where vision ends: Generalizable policies for contact-rich manipulation," *3rd RSS Workshop on Dexterous Manipulation: Learning and Control with Diverse Data*, 2025.

[36] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Contact-rich manipulation of a flexible object based on deep predictive learning using vision and tactility," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 5375–5381.

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

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

[39] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.

[40] W. Tong, J. Mi, X. Yi, N. Fazeli, and X. Huang, "Scalable open-source visuotactile sensor for 6-axis contact wrench estimation in tensegrity robots," *arXiv preprint arXiv:2607.15633*, 2026.