IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Visual-Force-Tactile Fusion for Gentle Intricate Insertion Tasks

Piaopiao Jin[1], Bidan Huang[2†], Wang Wei Lee[2], Tiefeng Li[1], and Wei Yang[1]

***Abstract*—This paper proposes a new approach for improving robotic manipulation tasks that require both precision and high compliance through multisensory fusion. By integrating visual, force, and tactile feedback, our approach enhances performance in delicate insertion tasks. We introduce a unified framework that combines information from these sensors to guide the entire manipulation strategies. Experiments in simulated and physical environments demonstrate that our method outperforms traditional single and dual-modality approaches regarding precision, gentle interactions, and robustness. We also provide a detailed analysis of the results to examine the role of each modality during manipulation. The experiment videos are available at https://sites.google.com/view/vft-fusion-insertion.**

***Index Terms*—Sensor Fusion, Force and Tactile Sensing, Machine Learning for Robot Control**

I. INTRODUCTION

AUTOMATING contact-rich tasks in robotics is crucial, especially when they are risky or time-consuming for humans. Tasks such as inserting pegs with uncertain or complex shapes require accuracy, compliance, and adaptability. Humans achieve these tasks by combining vision and haptic senses (force and tactile). Vision helps identify object geometries and positions, while haptic feedback provides insights into environmental interactions. Therefore, it is believed that by utilizing similar sensory modalities, robots can attain human-like dexterity. To this end, camera, force/torque sensor, and tactile sensor have been developed and used to guide the robotic automatic manipulation.

The integration of visual and haptic[1] perception in manipulation tasks has been extensively studied, especially with machine learning techniques. Applications include object recognition, slip detection, and pose estimation [1], [2], [3]. These algorithms primarily focus on visual-haptic perceiving and understanding. To enable autonomous robotic interaction with the environment, it is crucial to extract actionable information from the visual-haptic sensors and generate appropriate actions [4], [5]. In [6], [7], the researchers have integrated visual and force/torque data for peg-in-hole tasks.

Manuscript received: Oct. 30, 2023; Revised Feb. 4, 2024; Accepted March 12, 2024.

This paper was recommended for publication by Editor Jens Kober upon evaluation of the Associate Editor and Reviewers' comments.

† denotes the corresponding author.

[1] P. Jin, T. Li, and W. Yang are with the Center for X-Mechanics, Department of Engineering Mechanics, Zhejiang University. This study was conducted during P. Jin's internship at Tencent Robotics X. {piaopiaojin, litiefeng, yangw}@zju.edu.cn

[2] B. Huang and W. Lee are with Tencent Robotics X. {bidanhuang, wwlee}@tencent.com

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

[1]Haptics refers to the science and technology associated with the sense of touch. Haptics in this paper include the force/torque data from the six-axis force/torque sensor and the tactile data from the tactile sensor.

Different from the force/torque sensor that reflects the robot-environment interaction force, the tactile sensor [8], [9], [10] provides detailed contact information, such as pressure distribution between the manipulated object and the gripper. A number of researchers have developed visual-tactile algorithms for manipulation tasks. For example, Li *et al.* [11] propose a multisensory self-attention model that combines visual, acoustic, and tactile data for robotic manipulation tasks. Similarly, Hansen *et al.* [12] fuse visual and tactile perception for simulated tasks. While these works demonstrate impressive performance in visual-haptic fusion and object manipulation, the challenge of fusing visual-haptic multimodality for precise and gentle contact-rich tasks remains open.

In this paper, we present a novel approach to fusing visual, force, and tactile inputs for contact-rich manipulation tasks. Our method begins by developing modality-specific encoders to capture the distinct features of each sensor. We then fuse these multiple sensory inputs using curriculum policy learning, which organizes the learning process based on task difficulty, modality inputs, and manipulation gentleness requirements. Each modality provides distinct information, and their fusion results in a more comprehensive and precise representation of the environment and manipulated objects. By leveraging the complementarity of multiple sensory inputs, our algorithm learns an end-to-end manipulation strategy, resulting in actions that are both more precise and compliant compared to using a single modality.

Our method is validated on a series of insertion tasks in simulated environments [13] as shown in Fig. 1. Moreover, our system is also robust to multi-step long-horizon tasks. The direct sim-to-real transfer of the multimodal perception and control system also demonstrates the practicability of the multimodal algorithm on the physical environments. The contributions of this work can be summarized as follows:

- We have developed a multisensor fusion approach that effectively cooperates visual, force, and tactile data to guide gentle and precise insertion processes.
- The proposed approach has been validated through a series of intricate insertion tasks in both simulated and real-world environments, demonstrating its effectiveness.
- We have introduced the gentle interaction requirement to better exploit the complementarity of multimodality and enhance gentle insertion performance.
- We have conducted a comprehensive analysis of the multisensor fusion results and examined the individual contributions of each sensor during manipulation.

This study is an extension of our previous work [7] that emphasizes the integration of tactile perception, demonstrating the potential of incorporating additional modalities into

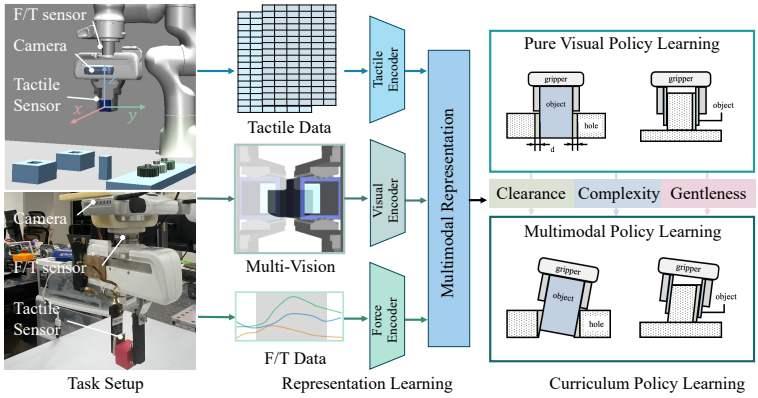

Fig. 1. Overview of the visual-force-tactile multimodality fusion and policy learning methodology. This control framework includes the sequential stages of multimodal data collection, multimodal data processing and fusion, and multimodal curriculum policy learning.

manipulation tasks. The improved performance in gentle and precise insertion tasks highlights the benefits of leveraging multiple modalities to achieve more effective and efficient manipulation. The comprehensive analysis of this study offers valuable insights for future research in this field.

II. RELATED WORK

A. Insertion Tasks

Peg-in-hole insertion tasks have been intensively studied for their close relevance to the manufacturing industry. The basic setup typically involves fixed round pegs overlapping their corresponding holes. To address the positional mismatches between pegs and holes, researchers have analyzed the relationship between peg geometry and force signals. Impedance or force control strategies are frequently employed in these methods, as referenced in [14], [15]. As computer vision techniques have advanced, robotic systems have become capable of managing more intricate scenarios, thereby reducing dependence on initial configurations [16]. Typically, the algorithm implements force control along the constraint axis while employing visual servoing to regulate motion in other dimensions. Recent advancements have further broadened the scope of insertion systems to encompass tasks like electronic component placement and intricate gear assembly [17], [18]. However, these methods are confined to scenarios wherein pegs are inserted from a top-down orientation and assume that the roll and pitch mismatches are zero. Although there are some works achieving 6-DoF peg insertion based on 3D visual feedback, the clearance is rather large (4 mm) [19].

Our study tackles challenges unaddressed by the aforementioned paradigms. Specifically, we explore peg-hole insertion tasks under exceptionally narrow clearances (up to 0.1mm) and gentle interaction requirements. Moreover, we also explore the insertion tasks where insertion directions are not directly observable but are explored through interactions.

B. Visual-haptic Perception and Control for Robots

Visual perception extracts diverse object and environmental attributes [20]. Consequently, many robotic systems rely solely on visual perception to close the control loop. However, the loss of the touch sensors greatly restricts the system's ability to exploit the environment in contact-rich scenarios. To enable better interaction and manipulation performance, some approaches equip the robot with touch sensors, namely the six-axis force/torque (F/T) sensor and the tactile sensor. The F/T sensor captures the interaction force between the robot and the environment and is often employed in compliant controllers [21], [22]. Unlike force perception, tactile data reflects contacts between the manipulated object and the end-effector, thus providing a more detailed view of their interaction. Driven by advances in haptic technology and a growing recognition of the synergistic relationship between vision and touch [23], [24], [25], the interests in visual-haptic perception and control have surged. Taunyazov *et al.* introduce an event-based tactile sensor alongside an event-driven perception system, which is applied to two robotic tasks: container classification and rotational slip detection [23]. Beyond combining visual and tactile data directly, research has also explored the reliability of each modality at different stages and has segmented tasks into separate phases [26].

Although the above studies have explored the fusion of vision and touch, the integration of visual, force, and tactile perception has rarely been considered. Our work aims to substantiate that unifying the three sensor modalities holds the potential to enhance efficiency and performance within intricate insertion tasks.

C. Reinforcement learning-based Contact-rich Manipulation

Reinforcement learning has found broad application in manipulation tasks that enable agents to acquire skills automatically through interactions. The training and deployment of the reinforcement learning algorithms on real machines are usually impractical due to their sample inefficiency. This is especially problematic for contact-rich tasks where extensive interactions are required. Some researchers have turned to model-based methods, which can bypass the need for extensive exploration [22]. Others train policies in simulation and then transfer them to the real world using domain randomization to bridge the gap between the two environments [21], [27]. However, the transfer of the multimodal policy from simulation to reality has been rarely reported. In our work, the visual-force-tactile perception and control system can be transferred from simulation to reality without additional training.

III. TASK DESCRIPTION AND METHOD OVERVIEW

The objective of this paper is to develop a framework for multimodal perception and manipulation. We verify this framework with a series of insertion tasks.

Task 1: The robot inserts a peg into a vertically oriented target hole, which has a 0.1 mm clearance and a 50 mm depth.

Task 2: In a reverse configuration, the robot positions a ring over a vertically standing peg under the same requirement.

Task 3: The robot conducts an angled peg insertion into a slanted hole with 0.5 mm clearance and a 50 mm depth.

Task 4: In a reverse configuration, the robot inserts a ring into an irregular peg under the same requirement.

Multi-step Insertion Tasks: The robot sequentially grasps and positions various objects like gears, pegs, and rings of 0.1 mm clearance into their respective target locations.

In intricate insertion tasks, the robot exhibits different dynamical properties before and during contact. Prior to contact, the robot moves in free space. Upon contact, the robot moves under environmental constraints, where even small displacements can result in significant contact force or task failure. Correspondingly, modality perception demonstrates distinctive attributes during these two phases. Visual modality operates in both phases, while force/tactile data only function upon contact. To this end, researchers typically break down insertion tasks into two phases. First, they use a visual detector to localize the target before making any contact. Then, they rely on touch sensors to guarantee a successful insertion during the contact. While this two-phase approach is straightforward, it requires the robot to switch between two controllers, which can cause instability in the system.

In this paper, we propose a single-phase approach that directly maps multimodal perception, i.e., vision, force/torque data, and tactile data to the robot's movements during the entire process. Precisely, visual perception is facilitated by two Realsense d435 cameras, force perception is derived from the ATI MINI40 six-axis force/torque sensor, and tactile perception is obtained via the gripper tactile sensor (piezoresistive) developed in our lab [9], as illustrated in Fig. 1. The robot motions include the incremental position and orientation commands $< \Delta X, \Delta q >$. To efficiently map the multimodal perception to the motion commands, we first propose modality-specific encoders to convert the raw visual, force, and tactile sensory data to the features v_{vision}, $v_{force}, v_{tactile}$. Next, the multimodal features are developed based on the concatenation of the respective features ($v_{vision} \oplus v_{force} \oplus v_{tactile}$). Subsequently, we propose curriculum multimodal policy learning to map the multimodal features ($v_{vision} \oplus v_{force} \oplus v_{tactile}$) to the incremental motion command $< \Delta X, \Delta q >$. The motion command is then executed by a compliant motion controller proposed in [28]. The overview of the perceptual and control framework is illustrated in Fig. 1.

IV. METHOD

A. Multimodal Perception and Representation Learning

1) Visual Perception: The relative pose between the manipulated object and the target is characterized by six parameters: E_x, E_y, E_z, E_ϕ, E_ρ, and E_θ, with each representing translational, roll, pitch, and yaw mismatches. Previous works mainly focus on the situation where the roll and pitch mismatches are zero, this paper considers that the pitch and yaw mismatches are zero. This task setup is more challenging because the roll mismatch E_ϕ is not directly observable from vision.

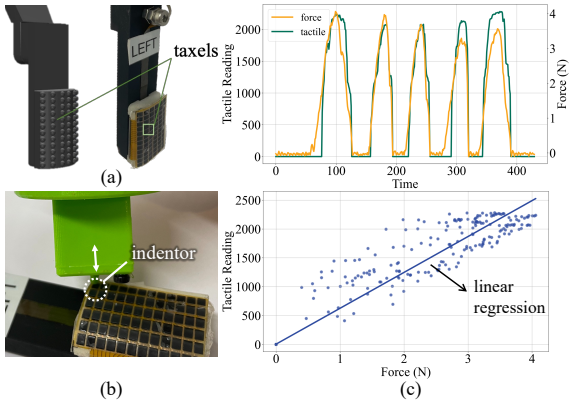

Fig. 2. (a) The simulated (left) and physical (right) tactile sensors. (b) The tactile calibration process with the indentor exerting/releasing force on a taxel. (c) The top row is the tactile-force readings during the calibration process on one taxel. The bottom row shows the linear regression on the collected tactile-force data from the top row.

Leveraging the front and rear in-hand cameras, we develop a self-supervised neural network that predicts solely two Booleans corresponding to E_x and E_y. The network predicts whether they are positive or negative. The two RGB images undergo separate processing via ResNet50 [29] backbone networks, reducing them to a 128-dimensional feature vector (v_{vision}). The visual feature is then input to a three-layer multi-layer perceptron (MLP), predicting the spatial alignment between the object and the target location. To train the self-supervised visual encoder, we use a binary-class cross-entropy loss with Adam optimizer. We train the network for 20 epochs with batch size 32 and learning rate $1e^{-4}$.

2) Force Perception: A six-axis force/torque sensor is mounted between the robot's flange plate and the gripper. It captures the interaction force and torque between the robot and the environment, which is represented as $f_{raw} = [f_x, f_y, f_z, \tau_x, \tau_y, \tau_z]$. Because this experiment doesn't consider yaw rotational mismatch, τ_z is irrelevant to the task and is disregarded in the perception process. For force perception, the most recent 5 readings from the six-axis F/T sensor are considered, creating a 5 $\times$ 5 temporal reading matrix that is subsequently flattened into a 25-d vector (v_{force}). The temporal information reduces the data noise and better reflects the dynamic interactions compared to the instant data.

3) Tactile Perception: This experiment uses two resistive tactile sensors [9] affixed to the parallel grippers. As illustrated in Fig. 2(a), the tactile sensor array, which comprises a 12 $\times$ 6 grid of taxels (sensing units), is bent into a curved surface. When functioning, every taxel generates a voltage response proportionate to the normal pressure exerted upon it. Subsequently, this voltage is converted into a tactile signal operating at a frequency of 100 Hz. To facilitate experimentation, we construct a simulated model of the tactile sensor in MuJoCo. Each taxel is modeled as a touch sensor to measure the contact signal proportional to the applied normal force (Fig. 2(a)).

To extract task-relevant tactile signals and filter out irrelevant data, we define the tactile observation tac_{ijt} by deducting the initial reading tac_{ij0}. In this context, i and j denote the row and column taxel indexes, and t is the timestep. This modification ensures that the tactile observation is sensitive to changes stemming from interactions while remaining unaf-

fected by initial grasp configurations. What's more, as depicted in Fig. 2(a), the tactile sensor features a grid of 6 columns. The curved topography of the finger surface makes the edges of taxels challenging to activate during parallel object grasping. Consequently, the first and last columns of the tactile arrays show limited responsiveness in our experiments. To address this, we average the readings from the first two columns and the last two columns into two distinct new columns. To further reduce the tactile dimensionality, the data from the first and second rows are averaged, as are the third and fourth rows, the fifth and sixth rows, the seventh and eighth rows, the ninth and tenth rows, and the eleventh and twelfth rows. Consequently, the initial 12×6 tactile sensor matrix is condensed into a 6×4 matrix. The above process could be seen as the conduction of mean pooling with irregular kernel and stride sizes as shown in Eqn. 1. m and n are the indexes to iterate the pooling window. $2i$ and $S(j)$ represent the start row and column indexes of the pooling window. 2 and $R(j)$ represent the number of rows and columns in the pooling window. Specifically, i iterates from 0 to 5, and j from 0 to 3. $R(j)$ is 1 when j equals to 1 or 2, and is 2 under other conditions. $S(j)$ is j when j equals to 0, and is $j+1$ under other conditions. After flatting and normalization, the tactile representation for two fingers is a 48-d feature vector ($v_{tactile}$).

$$tac^{'}[ij] = \frac{1}{2R(j)} \sum_{m=0}^{1} \sum_{n=0}^{R(j)-1} tac[2i+m, S(j)+n] \tag{1}$$

B. Curriculum Multimodal Fusion and Policy Learning

Our goal is to endow robots with the capability to perform a series of intricate insertion tasks by leveraging visual, force, and tactile perception. To learn the multimodal manipulation strategy, we model the insertion task as a finite-horizon, discounted Markov Decision Process (MDP) $\mathcal{M}$, which has a state space $\mathcal{S}$, an action space $\mathcal{A}$, the state transition dynamics $\mathcal{T}$:$\mathcal{S} \times \mathcal{A} \to \mathcal{S}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a horizon $\mathcal{T}$, and a discount factor $\gamma \in (0, 1]$. To determine the optimal stochastic policy $\pi : \mathcal{S} \to \mathbb{P}(\mathcal{A})$, we maximize the expected value of the discounted reward. The following are the specifics of the observation, action, reward, training techniques, and algorithm details employed in reinforcement learning.

1) observation space: To assess the benefits of the visual-force-tactile perception system, we conduct ablation studies on the following models, each using different sensory inputs:

- *visual model*: This model solely employs 128-d visual features for perception (v_{vision}).
- *visual-force model*: The observation space of the model comprises 153 dimensions, incorporating 128-d visual features and 25-d force features ($v_{vision} \oplus v_{force}$).
- *visual-tactile model*: The visual-tactile model operates within an observation space of 176 dimensions, integrating 128 dimensions for visual features and 48 dimensions for tactile features ($v_{vision} \oplus v_{tactile}$).
- *visual-force-tactile model*: The proposed method integrates visual, force, and tactile data to form a 201-d observation input ($v_{vision} \oplus v_{force} \oplus v_{tactile}$).

2) action space: The action space for task 1, task 2, and the multi-step insertion tasks consists of three dimensions (ΔX = [Δx, Δy, Δz]), signifying the end-effector's incremental displacements along the x, y, and z axes. The action space for task 3, and task 4 is a 4-d vector $< \Delta X, \Delta q >$. It represents the incremental displacements along the x, y, and z axes, as well as an incremental roll (ΔX = [Δx, Δy, Δz], Δq = $\Delta \gamma$).

3) curriculum policy learning procedure: Learning the visual-force-tactile manipulation policy directly for intricate insertion tasks encounters two significant challenges: the inherent difficulty of the tasks themselves and the high dimensionality of multimodal features. To address these challenges and expedite the learning process, we adopt a two-stage curriculum-based approach. **Pure visual policy learning**: In the initial stage, we focus on learning a pure visual policy for comparatively less intricate tasks. Tasks 1 and 2, distinguished by wider clearance and a simplified manipulation action space (ΔX = [Δx, Δy, Δz]), serve as our initial training scenarios. **Continual learning with multimodal fusion**: In the second stage, we introduce haptics into the learning process and tackle more challenging tasks. These tasks include reduced clearance in tasks 1 and 2 and increased task complexity in tasks 3 and 4. Furthermore, the action space expands to include (ΔX = [Δx, Δy, Δz], Δq = $\Delta \gamma$). To this end, the input and output layers of the policy network in RL are modified accordingly to accommodate the additional modalities and higher dimensional actions. The primary objective in the first stage is to establish a basic control strategy solely from vision. As the manipulation intricacy amplifies in the second stage, the system relies more on the combination of visual, force/torque, and tactile information to better explore and manipulate.

4) reward function design: The reward function is discrete and based on the insertion depth. A reward of 0.25 is provided at the manipulated object first attains 25 %, 50 %, 75 %, and 100 % of the total depth in the pure visual policy learning phase. Conversely, if the peg slips or tilts from the gripper, the agent terminates the episode and receives a penalty of -0.2. In the **continual learning with multimodal fusion** stage, the reward function is expanded with the gentle interaction requirement. Specifically, if the interaction force is larger than a threshold of 10 N, the episode is terminated with a penalty of -0.2. The reward function can be expressed as in Eqn. 2, where D denotes the total insertion depth.

$$r_t = \begin{cases} 0.25 \ \ if \ depth \in [0.25D, 0.5D, 0.75D, 1.0D] \\ -0.2 \ \ if \ early \ termination \\ 0 \quad otherwise \end{cases} \tag{2}$$

5) Learning Algorithm Details: We utilize Proximal Policy Optimization (PPO) [30] algorithm implemented by Stable-Baselines3 [31]. The policy network is a three-layer MLP, whose hidden layers have 32 neurons.

C. Sim-to-real Transfer Techniques

Simulated and physical environments differ significantly in terms of perception and control. Directly implementing the simulated system on a real machine is not feasible. To bridge the reality gap, multiple efforts are made.

1) visual domain randomization: A notable discrepancy between the simulated and physical environments is the significant variation in lighting conditions, object textures, and materials. Moreover, the resolution, quality, and field of view

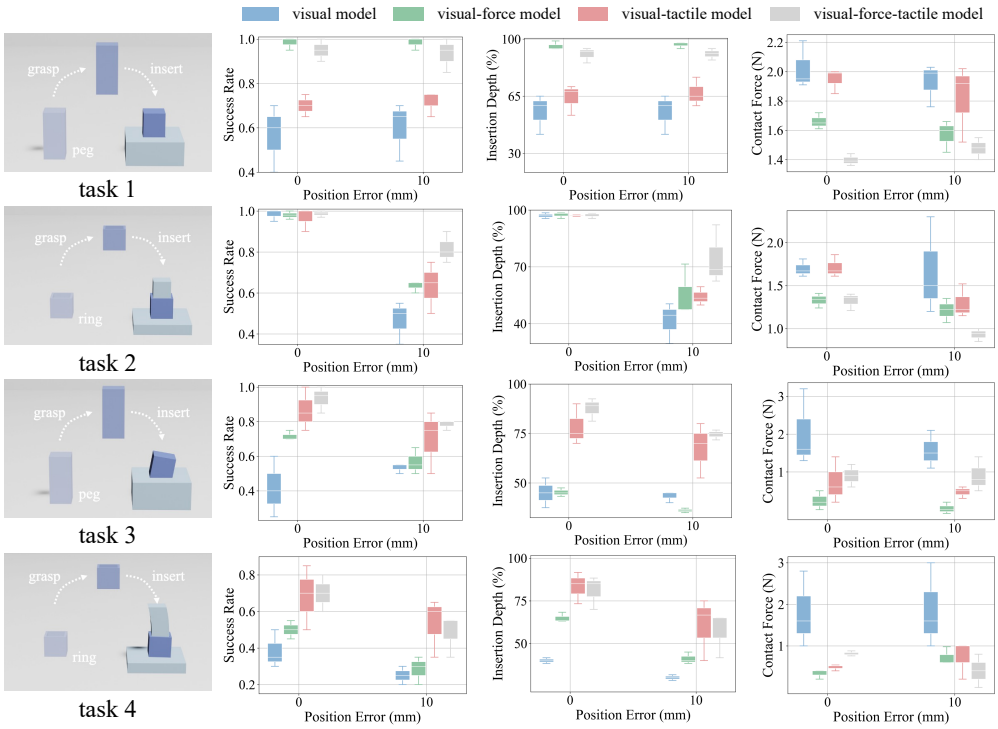

Fig. 3. Performance comparisons of the four multimodal systems across various insertion tasks, where the robot grasped and transported the object before executing the insertion policy. Two groups of initial conditions are considered, with position errors (distance between pegs and holes in the $x-y$ plane) around 0 mm and 10 mm. For each task, the policy was initialized with three random seeds and each seed was tested 20 times. For each seed, the success rate, insertion depth, and contact force were averaged. The bars in the graph represent the medium, minimal, and maximal data of these averages.

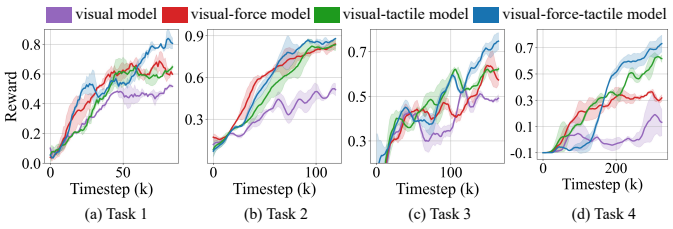

Fig. 4. Training episode rewards of the modality ablative models with three random seeds.

of the images in the simulated and physical platforms do not align. To make the visual features and policies developed in simulation adaptable for transfer, we employ a series of domain randomization techniques. Initially, we randomize lighting, colors, and textures within the simulations. Then, we apply Gaussian blurring, introduce random shadows, and add white noise to the rendered images.

2) tactile calibration: In addition to the reality gap in visual perception, differences exist in tactile perception between the simulated and physical environments. In the simulation, the tactile signal is proportional to the applied normal force, but the relationship between tactile and force data on the physical sensor remains uncertain. To establish this connection, we employ a series of calibration procedures within real-world scenarios. The initial step involves using a calibration indentor to tap and apply force to the tactile taxel. As depicted in Fig. 2(b), the indentor is securely attached to the F/T sensor, facilitating the recording of tactile-force data pairs. This tapping-release process is iterated five times per taxel, amassing adequate calibration data (as shown at the top of Fig. 2(c)). Subsequently, we approximate the tactile-force relationship as a linear function and execute linear regression using the captured data, as visualized in the bottom of Fig. 2(c). The tactile-force relationship could be approximated with $Real_{tac_{ij}} = k_{ij} * f_{ij} + b_{ij}$ where i and j denote the row and column index of the taxel. k_{ij} and b_{ij} are the approximated linear regression parameters. $Real_{tac_{ij}}$ is the tactile reading in the physical environment, and f_{ij} the force readings along the tap-release direction. For obtaining tactile observations in real scenarios, the raw tactile data $Real_{tac_{ijt}}$ at timestep t is processed as follows. Firstly, the current tactile data is subtracted by the initial tactile signal, which is $Real_{tac_{ijt}} - Real_{tact_{ij0}}$, and the mapped force observation is $\frac{Real_{tac_{ijt}} - Real_{tact_{ij0}}}{k_{ij}}$. Next, the data is normalized with the mean and the variance.

V. EXPERIMENTS DESIGN

The primary goal of our experiments is to examine the efficacy of the visual-force-tactile perception and control system in intricate insertion tasks. Specifically, the experiments are structured to address the following research questions (**RQ**s):

- RQ1. How does the performance of our current approach compare to the modality ablative models?
- RQ2. How does the gentleness requirement affect the insertion performance?
- RQ3. What are the contributions of the single modalities in the multimodal system?
- RQ4. How well do multimodal strategies perform on multi-step long horizon insertion tasks?
- RQ5. How effective is the sim-to-real methodology?

Evaluation Metrics: In all the tasks, a trial is considered successful when the manipulated object reaches a depth of 50 mm within 300 timesteps. Each timestep corresponds to a

TABLE I
SUCCESS RATE OF SIMULATED ROBOT EXPERIMENTS

Tasks	task 1	task 2	task 3	task 4	multi-ring	multi-gear	multi-peg
Trials	120	120	120	120	50	50	50
Results	94%	90%	86%	60%	90%	76%	90%

duration of 0.05 seconds in simulation time. Conversely, a trial is classified as unsuccessful if it doesn't meet these criteria.

VI. EXPERIMENT RESULTS

A. Ablative Models Performance Evaluations (RQ1)

To evaluate the effectiveness of the multimodal perception and control system in intricate insertion tasks, three key properties were examined, namely the success rate, insertion depth, and average contact force. For each task, we tested the four models' performance from two different categories of initial starts. Specifically, in the first category, the initial positional distances between the pegs and holes along the $x-y$ plane were around 0 mm. In the second category, the initial positional distances were around 10 mm. At each initial configuration, a total of 60 trials were conducted with three different random seeds. From Fig. 3, we could discover that the visual-force-tactile system exhibits an exceptional success rate nearing 100 % in both task 1 and task 2. Slight performance declines occur when the initial position errors are enlarged. In the more intricate scenarios of task 3 and task 4, where insertion directions are not observable, the visual-force-tactile model maintains a success rate exceeding 70 %, particularly when initial position errors are minimal.

In comparison to the ablative models-namely the *visual model*, *visual-force model*, and *visual-tactile model*-the proposed model demonstrated compelling advantages. Specifically, the *visual model*, while capable of task completion, recorded notably lower success rates, particularly under challenging conditions. Its insertion depth and interaction force profiles were notably inferior compared to the proposed model. The *visual-force model* achieved good performance in tasks demanding known insertion directions (task 1 and task 2), demonstrating high success rates, substantial insertion depths, and controlled contact forces. However, the *visual-force model* faced limitations in more intricate tasks (task 3 and task 4). On the contrary, the *visual-tactile model* surpassed the *visual-force model* in tasks requiring intricate contact interactions (task 3 and task 4) with a higher success rate and larger insertion depth. But the *visual-tactile model* had limited capabilities in task 1 and task 2. The performance difference between the *visual-force model* and *visual-tactile model* reflected the different characteristics of the force/torque and tactile data. The force/torque data that reflects the overall interactions between the robot and the surroundings is more valid in upright insertion tasks. While the tactile data that reflects the interactions between the gripper and the manipulated object is more valid in occluded and intricate scenarios. Although the *visual-force model* and the *visual-tactile model* exhibited their respective strengths, it is the fusion of visual, force, and tactile data in our proposed approach that led to the most promising performance across all four tasks. The learning curves visualized in Fig. 4 also validate that the fusion of

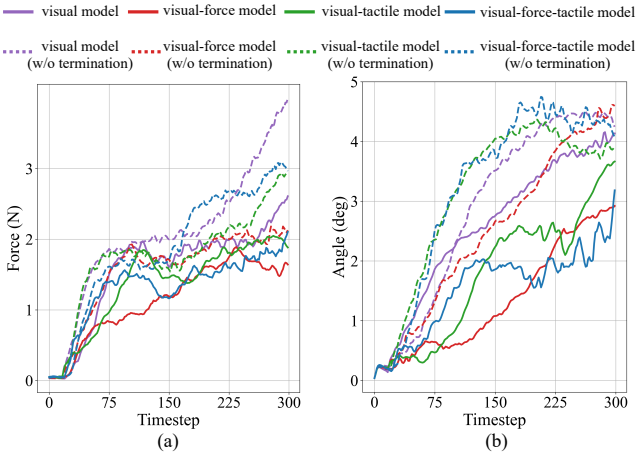

Fig. 5. Illustrations of the interaction force (a), and the relative rotation (b) between the object and the fingers during the insertion process. The results are averaged by 20 successful insertions in task 1.

the vision, force, and tactile data yields the most stable and promising results. The overall results of the visual-force-tactile system are presented in TABLE I.

B. Gentleness Requirement Efficiency (RQ2)

To evaluate how the gentle interaction requirement affects the insertion performance, we trained the multimodal policies in task 1 without the 10 N termination criterion and visualize the interaction force in Fig. 5(a). The manipulated object was not fixed, leading to its motions in the insertion process. We illustrate the rotations between the object and the fingers in Fig. 5(b). From Fig. 5(a) we have observed that the interaction force in the models without gentleness requirement exceeds the models with it. From Fig. 5(b) we have observed that models with visual and haptics input, i.e. the *visual-force*, *visual-tactile*, and *visual-force-tactile* models actively exploit the complementary between vision and haptics with the gentle requirement. Specifically, all the multimodal models exhibited similar rotation properties without the gentle interaction requirement, relying on the visual input and not exploring the functionality of haptics. The *visual model* demonstrated similar insertion properties with the gentleness requirement while in the *visual-force*, *visual-tactile*, and *visual-force-tactile* models, the relative rotations between the object and the fingers were significantly reduced. The reduced interaction force and object rotations demonstrate that the gentle interaction requirement enforced the agent to actively exploit the complementary between vision and haptics and enable more gentle insertion performance.

C. Modality Contribution Analysis (RQ3)

To systematically assess the individual contributions of each modality in the multimodal policies, we calculated sensitivity factors for visual, force, and tactile modality in the four tasks. Specifically, the policy network can be represented as: $a_t = \pi_{mlp}(v_{t_{vision}}, v_{t_{force}}, v_{t_{tactile}})$ Here, $v_{t_{vision}}, v_{t_{force}}, v_{t_{tactile}}$ correspond to the visual, force, and tactile features at time step t, while a_t represents the action at that step. To deduct the sensitivity factors for visual, force, and tactile modality, we first derived the partial differentials of a_{kt} (k is the index of the action) with respect to $v_{t_{vision}}$, $v_{t_{force}}$, and $v_{t_{tactile}}$. Next, by adding up the absolute value of the partial

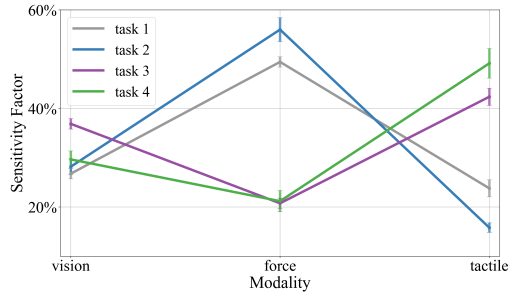

Fig. 6. Comparative analysis of the sensitivity factor for the three modalities across four insertion tasks.

differentials, we got the visual, force, and tactile sensitivity factors, denoted as S_{vision}, S_{force}, and $S_{tactile}$ respectively as shown in Eqn. 3, where N represents the dimension of the action a_t, and T the number of total timesteps. Fig. 6 represents the average modality sensitivity factors across 100 trajectories. Key observations can be discovered from this analysis. Firstly, the visual modality served as a foundational element for all four tasks, aligning with the understanding that vision provided comprehensive insights into environmental characteristics. This mirrored the human reliance on visual perception. Secondly, in tasks demanding precise manipulation but involving insertion directions orthogonal to the insertion plane (task 1 and task 2), the force modality provided more direct feedback than the tactile modality. The reason is that the force/torque sensor is more sensitive to the interactions between the robot and its environment, and provides more straightforward contact conditions in simpler insertion tasks.

$$
\begin{aligned}
S_{vision} &= \mathbb{E}_\tau \sum_{k=1}^{N}\sum_{t=1}^{T} \frac{1}{TN} \left| \frac{\partial a_{kt}}{\partial v_{t_{vision}}} \right| \\
S_{force} &= \mathbb{E}_\tau \sum_{k=1}^{N}\sum_{t=1}^{T} \frac{1}{TN} \left| \frac{\partial a_{kt}}{\partial v_{t_{force}}} \right| \\
S_{tactile} &= \mathbb{E}_\tau \sum_{k=1}^{N}\sum_{t=1}^{T} \frac{1}{TN} \left| \frac{\partial a_{kt}}{\partial v_{t_{tactile}}} \right|
\end{aligned}
\tag{3}
$$

However, in task 3 and task 4, where the insertion direction was not visually observable, the manipulation strategy relied more heavily on tactile feedback rather than force feedback. The tactile sensor is more responsive to the interactions between the object and the gripper, which is prominent in intricate tasks with unobservable insertion directions. In conclusion, due to their distinct responses to external stimuli, force and tactile perceptions played divergent roles in complex insertion tasks. Integrating vision, force, and tactile data equipped the robot with an improved ability to comprehend and leverage the intricate physics of unstructured environments. This fusion provided the robot with superior adaptability, enabling more efficient and successful object manipulation.

D. Multi-step Task Generalization Test (RQ4)

To assess the effectiveness of the multimodal system in multi-step tasks, we conducted tests in three different scenarios as shown in Fig. 7. These tasks required the robot first to grasp an object, transport it above the target location, and subsequently perform an insertion. After the insertion was completed, the robot released the gripper and approached the next object. The first task (upper row of Fig. 7) involves inserting and stacking three square rings. The second task

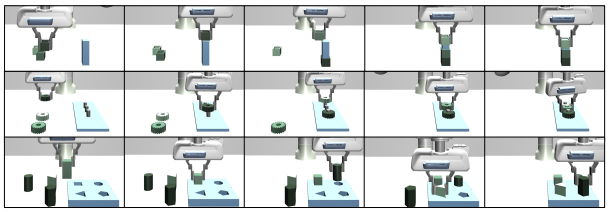

Fig. 7. Sequential snapshots of the robot executing a range of multi-step tasks. **Top:** The robot sequentially inserts and stacks three square holes. **Middle:** The robot inserts multiple gears. **Bottom:** The robot inserts multiple pegs into their matching holes.

TABLE II
SUCCESS RATE OF PHYSICAL EXPERIMENTS

tasks clearance	task 1 0.55 mm	task 2 0.61 mm	task 3 0.98 mm	task 4 1.15 mm	generalization 0.54 mm
visual model	3/10	3/10	0/10	0/10	17/40
visual-force model	6/10	5/10	4/10	0/10	24/40
visual-tactile model	7/10	5/10	8/10	2/10	31/40
visual-force-tactile model	7/10	8/10	4/10	2/10	29/40

(middle row) focuses on fitting multiple gears, and the third task (bottom row) centers on inserting multiple pegs into their matching holes. The multi-step tasks put a high demand for the efficiency and robustness of the system compared to the single tasks.

To evaluate the multi-step manipulation performance of the system, we conducted 50 trials on each task. In the first multi-step task, the robot could successively stack three square holes with a success rate of 90 %. Even though the multi-step gear insertion task was more challenging with occluded camera views, the robot could still succeed in 38 trials. As for the multi-step shape insertion task, the robot could sequentially insert the four shapes in 45 tests. The results demonstrated the multimodal system's generalizing capability across long-horizon multi-step tasks. This robust capability affirmed the system's capacity to adapt to a range of intricate scenarios.

E. Real Robot Experiments (RQ5)

To evaluate the proposed models' capacity for real-world deployment, we conducted the physical experiments in tasks 1-4 and the generalization test on pegs of different colors and shapes, namely the square, pentagon, triangle, and cylinder (Fig. 8). We conducted 10 trials for each modality model in tasks 1-4 and 10 trials for each shape in the generalization test. From TABLE II we can see that the results of the real robot experiments consist with the results of the simulation: policies trained with multimodal perception have a higher success rate than with single modality, while policies with tactile perception outperform those without. Particularly, the *visual model* showed the most limited capability across the four tasks. The *visual-force* model could not only capture the visual information but also the interaction force, showing enhanced capabilities and succeeding in task 3. Different from the force/torque sensor, whose readings were coupled with the motion of the end-effector and noisy in the motion of the robot, the tactile sensor was able to capture the direct interactions between the object, and the motion was more stable. The *visual-tactile* model showed better performance across tasks 1-4 and the generalization test. The *visual-force-tactile* model was expected to have the best performance across the four tasks, but its performance on task 3 was not as good as the *visual-tactile* model. It's presumed that the roll rotation of

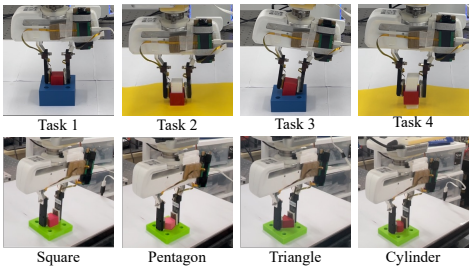

Fig. 8. Illustration of the physical experiments on tasks 1-4 and the generalization test on square, pentagon, triangle, and cylinder.

the end-effector introduced extra force noise which led to the unstable performance of the *visual-force-tactile model*.

Across the four tasks and the generalization test, two typical instances of task failure were observed, namely the object slippery and the object stuck in the hole. The major reason comes from the limitations of the tactile sensor used. Specifically, the tactile sensor exclusively captures normal forces, thereby omitting tangential forces. As a result, the relative motion between the gripper and peg, which plays a pivotal role in insertion tasks, remains inadequately monitored.

VII. DISCUSSION AND CONCLUSION

This paper is the first to integrate visual, force, and tactile perception for intricate insertion tasks. We demonstrate that our multimodal system could take advantage of each single modality and make the best potential of each modality. The compact perception and control system enables higher manipulation precision, gentler interactions, and generalization ability over multi-stage tasks. We expect our developed multimodal system from global vision, force/torque, and local tactile data to form a layered perceptual system more similar to a human perception and control system. Future work will focus on finding the best way to encode tactile data for contact-rich tasks and improving the reward function to achieve fast convergence in reinforcement learning.

REFERENCES

[1] 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, p. 3300–3307, Oct 2018.

[2] Z. Si and W. Yuan, "Taxim: An example-based simulation model for gelsight tactile sensors," *IEEE Robotics and Automation Letters*, 2022.

[3] W. Zheng, H. Liu, and F. Sun, "Lifelong visual-tactile cross-modal learning for robotic material perception," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1192–1203, 2021.

[4] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[5] N. Saito, T. Shimizu, T. Ogata, and S. Sugano, "Utilization of image/force/tactile sensor data for object-shape-oriented manipulation: Wiping objects with turning back motions and occlusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 968–975, 2022.

[6] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[7] P. Jin, Y. Lin, Y. Song, T. Li, and W. Yang, "Vision-force-fused curriculum learning for robotic contact-rich assembly tasks," *Frontiers in Neurorobotics*, vol. 17, 2023.

[8] N. Wettels, V. Santos, R. Johansson, and G. Loeb, "Biomimetic tactile sensor array," *Advanced Robotics*, vol. 22, pp. 829–849, 08 2008.

[9] Z. Ding, Y.-Y. Tsai, W. W. Lee, and B. Huang, "Sim-to-real transfer for robotic manipulation with tactile sensory," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6778–6785.

[10] S. Kim and A. Rodriguez, "Active extrinsic contact sensing: Application to general peg-in-hole insertion," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10 241–10 247.

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

[12] 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)*, 2022, pp. 8298–8304.

[13] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.

[14] T. Tang, H.-C. Lin, Y. Zhao, W. Chen, and M. Tomizuka, "Autonomous alignment of peg and hole by force/torque measurement for robotic assembly," in *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, 2016, pp. 162–167.

[15] M. H. Raibert and J. J. Craig, "Hybrid position/force control of manipulators," 1981.

[16] K. Hosoda, K. Igarashi, and M. Asada, "Adaptive hybrid visual servoing/force control in unknown environment," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS '96*, vol. 3, Nov 1996, pp. 1097–1103 vol.3.

[17] B. Tang, M. A. Lin, I. Akinola, A. Handa, G. S. Sukhatme, F. Ramos, D. Fox, and Y. Narang, "Industreal: Transferring contact-rich assembly tasks from simulation to reality," 2023.

[18] T. Davchev, K. S. Luck, M. Burke, F. Meier, S. Schaal, and S. Ramamoorthy, "Residual learning from demonstration: Adapting dmps for contact-rich manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4488–4495, 2022.

[19] B.-S. Lu, T.-I. Chen, H.-Y. Lee, and W. H. Hsu, "Cfvs: Coarse-to-fine visual servoing for 6-dof object-agnostic peg-in-hole assembly," 2023.

[20] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. A. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," *CoRR*, vol. abs/1803.09956, 2018.

[21] O. Spector and M. Zacksenhouse, "Learning contact-rich assembly skills using residual admittance policy," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 6023–6030.

[22] 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.

[23] T. Taunyazoz, W. Sng, H. H. See, B. Lim, J. Kuan, A. F. Ansari, B. Tee, and H. Soh, "Event-driven visual-tactile sensing and learning for robots," in *Proceedings of Robotics: Science and Systems*, July 2020.

[24] J. Kerr, H. Huang, A. Wilcox, R. Hoque, J. Ichnowski, R. Calandra, and K. Goldberg, "Self-supervised visuo-tactile pretraining to locate and follow garment features," 2023.

[25] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," 2019.

[26] T. Narita and O. Kroemer, "Policy blending and recombination for multimodal contact-rich tasks," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2721–2728, 2021.

[27] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1–8.

[28] Y. Lin, Z. Chen, and B. Yao, "Unified method for task-space motion/force/impedance control of manipulator with unknown contact reaction strategy," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1478–1485, 2022.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[31] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.