

# VTLA: Vision-Tactile-Language-Action Model with Preference Learning for Insertion Manipulation

Chaofan Zhang<sup>1,\*</sup>, Peng Hao<sup>2,\*</sup>, Xiaoge Cao<sup>1</sup>, Xiaoshuai Hao<sup>3</sup>, Shaowei Cui<sup>1,†</sup>, and Shuo Wang<sup>1</sup>

<sup>1</sup>State Key Laboratory of Multimodal Artificial Intelligence Systems,  
Institute of Automation, Chinese Academy of Sciences

<sup>2</sup>Samsung R&D Institute China-Beijing

<sup>3</sup>Beijing Academy of Artificial Intelligence

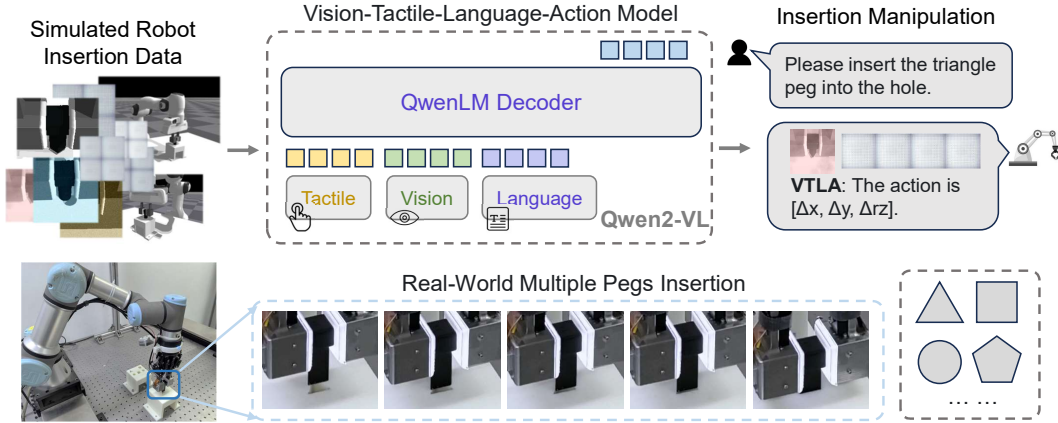


Figure 1: **Overview of VTLA.** The VTLA model learns a robotic manipulation policy integrated with vision, tactile, and language inputs from domain-randomized simulation data, enabling it to perform a variety of peg-in-hole tasks in the real world.

**Abstract:** While vision-language models have advanced significantly, their application in language-conditioned robotic manipulation is still underexplored, especially for contact-rich tasks that extend beyond visually dominant pick-and-place scenarios. To bridge this gap, we introduce Vision-Tactile-Language-Action (**VTLA**) model, a novel framework that enables robust policy generation in contact-intensive scenarios by effectively integrating visual and tactile inputs through cross-modal language grounding. A low-cost, multi-modal dataset has been constructed in a simulation environment, containing vision-tactile-action-instruction pairs specifically designed for the fingertip insertion task. Furthermore, we introduce Direct Preference Optimization (*DPO*) to offer regression-like supervision for the **VTLA** model, effectively bridging the gap between classification-based next token prediction loss and continuous robotic tasks. Experimental results show that the **VTLA** model outperforms traditional imitation learning methods (*e.g.*, diffusion policies) and existing multi-modal baselines (TLA/VLA), achieving over 90% success rates on unseen peg shapes. Finally, we conduct real-world peg-in-hole experiments to demonstrate the exceptional Sim2Real performance of the proposed **VTLA** model. For supplementary videos and results, please visit our project website: [VTLA](#).

**Keywords:** Contact-Rich Manipulation, Tactile Sensing, Large Language Model

## 1 Introduction

In contact-intensive manipulation tasks, such as precise object insertion [1, 2], vision is essential for environmental perception. However, humans naturally integrate tactile feedback to manage uncer-

\*Equal contribution

†Corresponding author: shaowei.cui@ia.ac.cn

tainties [3, 4, 5]. For example, aligning a key to a lockhole with an ambiguous position necessitates reliance on tactile feedback to compensate for limited visual information [6, 7, 8]. This highlights the importance of vision-tactile fusion in achieving robust robotic manipulation [9]. The recently emerging vision-tactile learning frameworks offer a promising foundation for robotic skill acquisition, allowing robots to interpret complementary contact-state signals to enhance grasping and manipulation [10, 11, 12]. However, existing studies primarily rely on training proprietary models on specific datasets, leading to challenges in generalization across diverse scenarios and a lack of adaptability for human-like extensive perceptual-motor reasoning [13, 14, 15].

Recently, Large Language Models (LLMs) have made significant advancements in human-like reasoning [16]. This progress has accelerated the development of Vision-Language-Action (VLA) models specifically designed for robotic manipulation [17, 18, 19, 20]. By aligning visual modalities with language modalities and leveraging LLMs for reasoning [21, 22, 23, 24], VLA models have significantly surpassed traditional imitation learning methods [25, 26] in their ability to generalize across diverse robotic platforms and task configurations [27, 28]. However, the lack of tactile feedback in most VLA systems limits their functionality to simple tasks, such as grasping and placing. This restriction hinders their applicability in contact-rich manipulation scenarios [29].

Tactile LLMs have demonstrated impressive understanding and reasoning capabilities in tactile perception, such as texture description and material recognition [30, 31, 32, 33, 34]. However, their application in language-conditioned action modeling remains nascent [35]. A Tactile-Language-Action (TLA) model [36] has been proposed for contact-rich manipulation tasks, demonstrating its potential in generalist tactile policy learning. Further investigation reveals two key limitations: 1) the tactile encoding and training/inference framework of the TLA model still have significant room for improvement; and 2) the absence of visual modalities imposes performance ceilings, as the robot lacks global perception capabilities.

To address these limitations, we introduce the Vision-Tactile-Language-Action (**VTLA**) framework, designed for effective contact-rich manipulation by integrating visual, tactile, and linguistic information. The technical contributions of VTLA are twofold. First, we design Vision-Guided Temporally Enhanced Tokens (VGTE) based on the VLM capabilities and the characteristics of vision-tactile manipulation tasks. By emphasizing visual tokens and enhancing temporal fusion before tokenization, VGTE mitigates the limitations of VLMs in temporal understanding and improves VTLA’s performance. Second, we employ Direct Preference Optimization (DPO) [37] to offer regression-like supervision for our action-conditional model. To further evaluate the Sim2Real transferability of **VTLA**, we created a vision-tactile robotic peg-in-hole assembly environment that replicates the simulation setup at a 1:1 scale. We then conducted real-world experiments on the VTLA model, which was trained exclusively on simulation data, across various peg-hole clearances and geometric configurations.

Our main contributions are summarized as follows:

- We propose **VTLA**, a novel vision-tactile fusion framework that integrates perception and language-action generation for contact-rich manipulation tasks.
- VGTE is designed to address the limited temporal reasoning capability of VLMs in vision-tactile manipulation. By emphasizing visual priors and incorporates temporal fusion prior to tokenization, VGTE enhances the cross-modal temporal reasoning of VTLA model.
- Preference Learning is introduced into VTLA to mitigate overfitting to ground-truth actions. By leveraging DPO to simulate regression-like supervision, VTLA gains richer training signals and performs enhanced generalization.
- Real-world insertion experiments show that **VTLA** outperforms current methods, demonstrating the effectiveness of the designed modules.

We hope this work will provide new insights for tactile-embedded VLA frameworks and inspire future research.

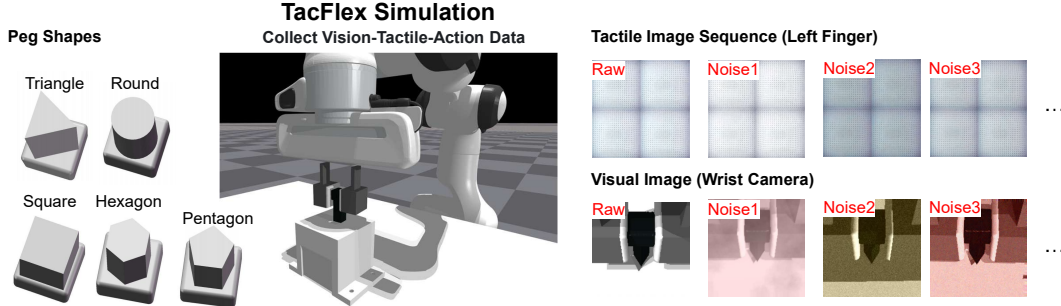


Figure 2: The data collection diagram and data examples of the VTLA dataset.

## 2 Related Works

**Vision-Tactile Learning** Vision-tactile fusion perception is crucial for improving robotic grasping and manipulation [38]. With the advent of deep learning, various fusion mechanisms—such as early feature fusion [39] and attention mechanisms [11]—have been explored for tasks like slip detection [40], grasp outcome prediction [10], and liquid pouring [12]. Reinforcement learning methods leveraging vision-tactile fusion have been proposed for peg-in-hole assembly [1, 2]. Recently, the combination of vision-tactile data acquisition with imitation learning has emerged as a promising area [41, 42, 43, 44, 45, 46]. However, these approaches often depend on specialized models trained on fixed datasets, limiting their out-of-domain generalization compared to general-purpose models.

**VLM for Robot Manipulation** Recent advancements in robotic manipulation showcase the potential of vision-language models (VLMs) with enhanced reasoning capabilities. RT-2 [15] formulates robotic actions as token sequences through fine-tuning on manipulation datasets, establishing the Vision-Language-Action (VLA) paradigm. Subsequent studies, such as RoboFlamingo [47] and OpenVLA [17], adopt similar methods. GR-1 [48] and GR-2 [49] employ a two-stage training approach, first pre-training on web-scale video corpora and then adapting to manipulation datasets. Innovations like RDT-1B [50] and PI [51] integrate stochastic modeling with diffusion objectives to enhance action sequence generation. However, existing VLA models mainly focus on visual modalities, limiting their use to relatively simple tasks like planar pushing and pick-and-place operations. This study investigates the integration of tactile perception into a vision-language foundation model, targeting contact-rich manipulation challenges such as peg-in-hole insertion.

**Tactile-Language Model in Robotics** Recent works in tactile-language model emphasize material understanding. Fu et al. [52] created a ChatGPT-assisted tactile-texture dataset using Tactile-Vision-Language Models, while Cheng et al. [31] expanded this and presented Touch100k. Despite these advancements, tactile data is still underutilized in robotic manipulation. Most recently, Jones et al. [35] embed tactile signals into VLA models for interaction policies. TLA [36] links tactile modalities and language to generate robot actions, indicating the potential of tactile modalities in language-conditioned robot learning. In this paper, we propose VTLA model to generate robotic manipulation policies by integrating visual, tactile, and language modalities.

## 3 Vision-Tactile-Language-Action Model

### 3.1 Data Collection

To enhance data collection efficiency and evaluate the Sim2Real generalization capability of the model, we adopt a synthetic data training approach followed by real-world validation. We construct a peg-in-hole assembly task in NVIDIA Isaac Gym with a self-built visuotactile simulator. This setup integrates a wrist-mounted camera and visuotactile sensors on the gripper fingertips to capture both visual and tactile observations during the assembly process, as illustrated in Fig. 2.

The assembly task is structured as follows: The gripper grasps a peg, positions it above the corresponding hole, and introduces a randomized 3-DOF misalignment in the x-axis, y-axis, and z-axis

rotation. The gripper then descends to attempt insertion. If a collision occurs, the attempt fails, and the gripper retracts for another try. If no collision happens before reaching the insertion depth, the task is successful. The maximum number of attempts is set to 15; otherwise, the task fails. In simulation, a randomized insertion strategy generates visual and tactile data, with task configurations and action labels aligned with those in TLA [36].

We collect data for five distinct peg-hole shapes with assembly clearances from 0.6 to 2.0 mm. The VTALA dataset comprises 28,000 assembly samples, each containing left/right tactile image sequences, a visual image, and an action label. The tactile image sequences are arranged in a  $2 \times 2$  grid, as shown in Fig. 2. To enhance zero-shot Sim2Real transfer performance, domain randomization techniques are employed during dataset generation to vary parameters in the simulation environment, task configurations, and both visual and tactile observations. Details on domain randomization are provided in the supplementary materials.

To facilitate training of the *VTALA* model, the dataset is organized in an instruction format. The `<|im_start|>` and `<|im_end|>` tokens mark the start and end of each dialogue round. Tactile and visual images are input sequentially with `<|vision_start|>` and `<|vision_end|>`. A text instruction specifies the task, including image types, peg shapes, and robot action requirements, while the action label serves as the ground truth. An example of *VTALA* data is provided below.

#### The Dataset Example of VTALA

```
<|im_start|>user
<|vision_start|> TactileLeftHand.png<|vision_end|> <|vision_start|> TactileRightHand.png<|vision_end|> <|vision_start|>
WristImage.png<|vision_end|> Given the tactile images from robot left and right fingertips and a wrist camera image during the
execution of peg-in-hole inserting tasks. Predict the correct robot action to insert the <Peg_Type> peg. Predict action: <|im_end|>
<|im_start|>assistant
[-0.9, 0.4, 0.013] <|im_end|>
```

### 3.2 Instruction Tuning with Vision-Guided Temporally Enhanced Tokens

**Vision-Guided Temporally Enhanced Tokens** While VLMs demonstrate superior generalization, their performance on specific tasks remains sensitive to the input prompts. This section presents the design of VTALA, which is tailored to the characteristics of both language models and visual-tactile manipulation tasks. VTALA introduces a vision-guided temporal enhancement mechanism to construct multimodal tokens that serve as temporally aligned inputs for the language model. This design enables more effective cross-modal reasoning by leveraging the multimodal understanding of pre-trained VLM, thereby facilitating superior performance in visual-tactile control task.

VTALA’s multimodal inputs focus on two key aspects: vision guidance and temporal enhancement. Prior studies [2] show that visual observations are crucial in the early stages of visual-tactile tasks. To reflect this priority and address the recency bias of language models [53], VTALA positions visual inputs after tactile inputs, bringing vision closer to action prediction. This strategy emphasizes the importance of visual information during initial manipulation phases, enabling VTALA to better utilize essential visual cues for contact-rich control.

Furthermore, VTALA employs temporally enhanced tactile inputs to improve the VLM’s reasoning of sequential data. While VLMs excel in cross-modal reasoning, they struggle with fine-grained temporal dependencies [54]. Although the Qwen-VL family uses 3D convolutions for video processing [55, 56], there is a domain gap between their high-level semantic tasks and the short-duration, low-level nature of tactile images in robotic manipulation. To bridge this gap, we encode tactile observations into image-like representations and extract temporally-aware tactile tokens using a Vision Transformer (ViT). This approach addresses the temporal reasoning limitations of VLMs and leverages their strengths in multimodal comprehension.

**Supervised Fine-Tuning** Guided by the above two designs, we fine-tune the VTALA model with the collected simulation dataset. Following prior works [17, 36], we formulate VTALA as a Next Token Prediction (NTP) task as shown in Fig. 3. Specifically, tactile and visual observations are first processed through a pre-trained vision encoder and a modality adapter to generate tactile and

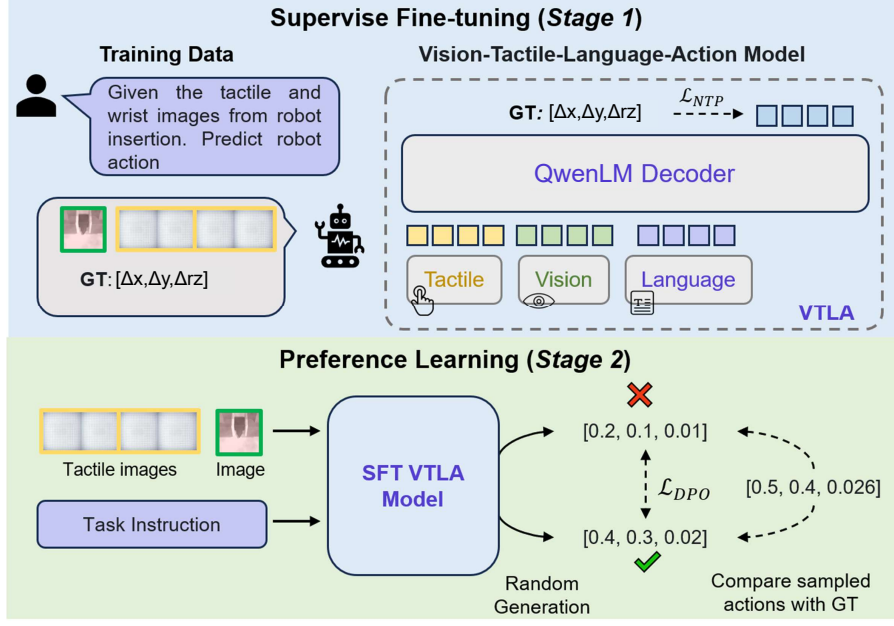


Figure 3: **The pipeline of VTLA.** In stage 1, the instruction dataset is created from simulation data using a vision-guided temporal enhancement, and the VTLA model is optimized with NTP loss. In stage 2, DPO is introduced to provide regression-like supervision, bridging the gap between VLM training and robotic continuous control, thereby enhancing performance.

vision tokens. Meanwhile, the textual instruction is tokenized to obtain text tokens. These generated tactile, vision, and text tokens then fed into a pre-trained LLM to predict the robot action. The entire model is fine-tuned using the NTP loss, defined as follows,

$$\mathcal{L}_{NTP} = - \sum_{n=1}^N \log P_{\theta}(x_n | x_{<n}), \quad (1)$$

where  $N$  denotes the total length of input tokens,  $x_n$  is the ground-truth token at position  $n$ , and  $x_{<n}$  represents the preceding tokens as context.  $\theta$  denotes the trainable parameters of the VTLA model. Following prior works [57], we freeze the parameters of the vision encoder and modality adapter, tuning only the language model to enhance performance.

### 3.3 Preference Learning

The instruction-tuned VTLA model shows limited performance due to a mismatch between robotic control tasks and the classification-based NTP loss. Robotic control is a regression problem that requires predicting continuous control signals, but Stage 1 of the VTLA model uses a classification-oriented NTP loss, overlooking this aspect. To address this, we reformulate the VTLA prediction task as a multi-label problem, enabling richer supervision through multi-label optimization. We specifically introduce Direct Preference Optimization [37] into VTLA, simulating a regression-like loss through preference learning.

The preference learning process is illustrated in Fig. 3. We first use the fine-tuned VTLA model to generate diverse action predictions from the same training samples. Based on their proximity to the ground truth, we create a preference dataset, labeling actions closer to the ground truth as chosen and others as rejected. Finally, we optimize the fine-tuned model on this dataset using DPO. The training objective is as follows:

$$\mathcal{L}_{DPO} = - \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_{\text{chosen}} | x)}{\pi_{\text{ref}}(y_{\text{chosen}} | x)} - \log \frac{\pi_{\theta}(y_{\text{rejected}} | x)}{\pi_{\text{ref}}(y_{\text{rejected}} | x)} \right), \quad (2)$$

where  $x$ ,  $y_{\text{chosen}}$ , and  $y_{\text{reject}}$  represent the input tokens, preferred responses, and rejected responses, respectively.  $\pi_{\theta}$  and  $\pi_{\text{ref}}$  are the trainable and frozen VTLA models, both initialized from the fine-



tuned model in stage 1. The scalar  $\beta > 0$  controls the preference signal’s sharpness, and  $\sigma(\cdot)$  is the sigmoid function. This loss function promotes a higher likelihood for the preferred response over the rejected one, aligning the output with preference-based supervision.

## 4 Experiments

We conduct a series of experiments to answer the following key research questions: **(Q1)** How does our proposed method compare with existing state-of-the-art approaches? **(Q2)** What is the impact of preference learning on model performance? **(Q3)** Can the proposed method generalize to real-world robotic applications?

### 4.1 Experiment Setup

**Existing and Ablation Methods** To address the above questions, we compare several baseline and ablation methods with the proposed VTLa on the dataset and various insertion tasks in simulation. 1) The *Diffusion Policy (DP)* [58] learns the insertion policy using vision and tactile observations. 2) A *Vision-Language-Action (VLA)* model is trained on wrist camera images from the dataset. 3) A *Tactile-Language-Action (TLA)* model is trained on tactile image sequences. 4) We tested the fine-tuned VTLa model under two generation configurations, obtaining 2,400 preference data points by comparing with the ground truth. Models trained with 1,000 and 2,400 preference data are denoted as *VTLa (w/ DPO-1k)* and *VTLa (w/ DPO-2k)*, respectively.

**Evaluation Metrics** For dataset evaluation, we define the Goal Convergence Rate (GCR) as the percentage of actions that are all correct in the  $x$ ,  $y$ , and  $rz$  directions. The L1 distance between output actions and action labels is used to evaluate performance in each direction. For the insertion task in simulation, we compute the success rate and average attempt steps for each method.

**Real-World Insertion Robot Setup** We use a 6-DoF UR3 robot arm with a Robotiq 2F-85 gripper for real-world insertion experiments. An Intel RealSense D405 camera is mounted on the wrist to capture vision images, while two GelStereos 2.0 sensors [59] on the gripper’s fingertips are used to obtain tactile observations during peg-hole collisions.

**Implementation Details** We utilized the LlamaFactory framework [60] for both SFT and DPO training. During the SFT stage, we used the Qwen2-VL 7B model as the base model, with a learning rate of  $5 \times 10^{-4}$ , a batch size of 64, and 10 epochs. In the DPO stage, the SFT model are served as the initialized model, with a learning rate of  $5 \times 10^{-6}$ , a batch size of 32, and 3 epochs.

### 4.2 Comparison with Baseline Methods

Table 1: Comparison of different methods on the dataset.

Method	ID				OOD			
	GCR(%)	L1 x	L1 y	L1 rz	GCR(%)	L1 x	L1 y	L1 rz
DP [58]	7.8	0.826	0.819	1.421	8.5	0.821	0.843	1.407
VLA [55]	46.1	0.210	0.247	<b>0.886</b>	29.5	0.353	0.351	1.221
TLA [36]	15.3	0.531	0.677	1.427	14.4	0.509	0.675	1.462
<b>VTLa (Ours)</b>	<b>47.3</b>	<b>0.181</b>	<b>0.224</b>	<u>0.904</u>	<b>31.2</b>	<b>0.305</b>	<b>0.324</b>	<b>1.136</b>

We compare the performance of VTLa with baseline methods on the dataset. We take 6k samples from the In-Distribution (ID) set, and 4k samples from the Out-Of-Distribution (OOD) set for evaluation, and the quantitative results are shown in Tab. 1. Furthermore, we evaluate these methods on various insertion tasks within the simulation environment. Specifically, square pegs with clearance of 2.0, 1.6, 1.0, and 0.6 mm are tested, following with pegs of various shapes with 0.6

Table 2: Comparison of different methods on the **square peg** insertion tasks with different clearances in simulation.

Method	2.0 mm		1.6 mm		1.0 mm		0.6 mm	
	Suc	Step	Suc	Step	Suc	Step	Suc	Step
DP [58]	42	2.47	32	2.63	28	4.85	22	3.54
VLA [55]	<b>100</b>	2.28	<b>98</b>	3.24	90	3.28	80	5.55
TLA [36]	94	3.27	90	3.60	80	4.97	80	5.48
<b>VTLa (Ours)</b>	<b>100</b>	2.12	<b>98</b>	2.87	<b>96</b>	4.64	<b>90</b>	5.91

mm clearance. Each insertion task is tested 50 times in simulation. The experimental results are presented in Tab. 2 and Tab. 3.

The experimental results in Tab. 1 and Tab. 2 show that the LLM-based models perform significantly better than the DP method in this insertion task (Tab. 2: 80+% vs 22% success rate in the 0.6 mm clearance insertion experiment). The VTLA and VLA model significantly outperform the TLA model in quantitative results (Tab. 1: GCR-ID, 46+% vs 15.3%). These findings highlight the key role of visual input in robot manipulation tasks. In addition, we find that VTLA with tactile encoding achieves better dataset performance (Tab. 1: GCR-ID, 47.3% vs 46.1%) and task success rate (Tab. 2: 0.6 mm clearance, 90% vs 80%) than VLA, which shows the effectiveness of temporally-aware tactile tokens in the proposed VTLA model.

In terms of generalization performance, the LLM-based models perform significantly better than the DP method (Tab. 3: OOD-Round, 90+% vs 10%). Furthermore, the VTLA model is better than TLA and VLA model in all dimensions of OOD data (Tab. 1: OOD-GCR, 31.2 vs 29.5/14.4).

In the pentagonal peg insertion task, VTLA achieves a higher insertion success rate (92% vs 82%/80%) and fewer steps (3.97 vs 4.41/4.60). Interestingly, all three models demonstrate comparable performance in the round peg insertion experiment, which may be attributed to the geometric isotropy of the round shape, making this task inherently less challenging.

Table 3: Comparison of different methods on the peg-in-hole task with 0.6 mm clearance in simulation.

Method	ID						OOD			
	Square	Triangle	Hexagon	Pentagon	Round		Suc	Step	Suc	Step
DP [58]	22	3.54	30	3.87	28	3.00	26	5.61	10	3.80
VLA [55]	80	5.55	82	5.02	84	3.83	82	4.41	94	4.81
TLA [36]	80	5.48	74	4.27	80	5.25	80	4.60	92	3.54
<b>VTLA (Ours)</b>	<b>90</b>	<b>5.91</b>	<b>88</b>	<b>4.53</b>	<b>90</b>	<b>4.68</b>	<b>92</b>	<b>3.97</b>	<b>92</b>	<b>4.74</b>

### 4.3 Ablation Study

Table 4: Ablation study on direct preference optimization.

Method	ID				OOD			
	GCR(%)	L1 x	L1 y	L1 rz	GCR(%)	L1 x	L1 y	L1 rz
VTLA (w/o DPO)	<b>47.5</b>	0.184	0.227	0.907	27.0	0.349	0.367	1.223
VTLA (w/ DPO-1k)	<b>47.5</b>	<b>0.181</b>	<b>0.224</b>	0.906	<b>31.4</b>	<b>0.305</b>	<b>0.324</b>	1.137
VTLA (w/ DPO-2k)	47.3	<b>0.181</b>	<b>0.224</b>	<b>0.904</b>	31.2	<b>0.305</b>	<b>0.324</b>	<b>1.136</b>

The ablation study on DPO is presented in Tab. 4. Experimental results show that preference learning with DPO significantly improves performance on both ID and OOD data. Specifically, VTLA-DPO-1k achieves a 16% improvement in GCR and approximately a 10% reduction in L1 error across all dimensions on OOD data. These improvements are attributed to the alignment of the DPO optimization objective with the nature of continuous robotic control tasks. By alleviating overfitting to sampled ground-truth actions during the SFT stage, DPO enhances the model’s generalization capabilities, particularly on OOD samples. Furthermore, we observe that increasing the size of the preference dataset does not lead to further performance gains. As the current preference data is generated by comparing outputs from only two sets of sampling configurations, we hypothesize that increasing diversity in preference data may be more effective than merely scaling dataset size.

### 4.4 Real-world Robotic Insertion

Table 5: Real-world insertion tasks on square pegs with different clearances.

	1.6 mm	1.0 mm	0.6 mm			
	Suc	Step	Suc	Step	Suc	Step
VTLA	100	1.60	100	1.95	95	4.31

Table 6: Real-world insertion tasks on different pegs with 0.6 mm clearances.

	ID						OOD			
	Square		Triangle		Hexagon		Pentagon		Round	
	Suc	Step	Suc	Step	Suc	Step	Suc	Step	Suc	Step
VTLA	95	4.31	95	3.94	95	3.52	100	1.85	100	5.2

In the real-world insertion experiment, we comprehensively evaluate the Sim2Real capabilities of VTLA and other LLM-based methods (TLA and VLA). Each insertion task is conducted 20 times

on the real robot. First, we test the ability of VTLA model, trained entirely in simulation, to handle varying assembly clearances in real-world insertion tasks. The results in Tab. 5 show that as the assembly clearance decreases, the average assembly steps of VTLA increases from 1.6 to 4.3. Despite this increased task difficulty, VTLA maintains a success rate above 95%. We further examine the generalization ability of VTLA to different peg-hole shapes in real scenarios. The results in Tab. 6 show that VTLA achieves a 100% assembly success rate on OOD peg shapes, even slightly outperforming its performance on ID peg shapes.

Compared with different LLM-based methods, the results in Tab. 7 show that both VTLA and VLA model with vision modalities achieve good performance (over 90%), but VTLA model with fusion of vision and tactile achieves better insertion efficiency (1.85 steps vs 2.3 steps in Tab. 7, Fig. 4-Case 1). Moreover, we find that the TLA model with only tactile observations faces a larger Sim2Real gap, with a success rate of only 30-40%, which is only half of that in simulation. As shown in the Case 2 of Fig. 4, under the same initial task states, VTLA completes the insertion within 3 steps, whereas TLA attempts 15 times but still fails to identify the correct direction, ultimately resulting in task failure. We suggest that exploring more advanced Sim2Real transfer methods may improve the real-world success rate of TLA model and potentially enable VTLA to handle more challenging insertion tasks.

Table 7: Comparison of different methods on the peg-in-hole task **with 0.6 mm clearance in the real world.**

Method	Triangle		Pentagon	
	Suc	Step	Suc	Step
VLA [55]	90	4.06	100	2.3
TLA [36]	30	2.00	40	1.88
<b>VTLA (Ours)</b>	<b>95</b>	<b>3.94</b>	<b>100</b>	<b>1.85</b>



Figure 4: Snapshots of real-world insertion using the proposed VTLA model and baseline methods (VLA and TLA). The initial state of the task is consistent, and the assembly clearance is 0.6 mm.

## 5 Conclusion

This study introduces the VTLA model, a vision-tactile-language-action model designed for contact-rich insertion manipulation tasks. The VTLA model is able to learn generalized visual-tactile skills based on language through a cross-modal fine-tuning process. Experimental results show that in the challenging fingertip peg-in-hole task, VTLA significantly outperforms traditional imitation learning methods and surpasses both TLA and VLA models, achieving an assembly success rate exceeding 90%. The VTLA model demonstrates strong task generalization capabilities across different assembly clearances and peg shapes. In addition, we also find that VTLA exhibits promising Sim2Real transfer performance, and the model trained using only simulated data can achieve an assembly success rate of 95% in the real-world insertion task.



**Limitations** Despite these encouraging results, VTLA has some areas for enhancement. One key aspect is aligning the tactile modality with the language modality, particularly regarding the semantic information of contact states for contact-rich manipulation tasks. In this paper, we utilize an off-the-shelf vision encoder to represent tactile inputs, which might lead to a loss of the unique features inherent in the tactile modality. Future work on dedicated tactile-language alignment is necessary. Another area for improvement is the fusion of visual and tactile modalities. While the deep integration of visual and tactile signals and features has been explored in proprietary models, corresponding research for LLMs remains limited. We believe this represents a promising topic for future work and encourage the community to explore these challenges together.

## References

- [1] 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)*, pages 8298–8304. IEEE, 2022.
- [2] 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*, 36(3):582–596, 2020.
- [3] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez. Tactile dexterity: Manipulation primitives with tactile feedback. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 8863–8869. IEEE, 2020.
- [4] Y. Wu, Z. Chen, F. Wu, L. Chen, L. Zhang, Z. Bing, A. Swikir, S. Haddadin, and A. Knoll. Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation. *arXiv preprint arXiv:2409.11047*, 2024.
- [5] J. Castaño-Amorós, I. d. L. Páez-Ubieta, P. Gil, and S. T. Puente. Visual-tactile manipulation to collect household waste in outdoor. *arXiv preprint arXiv:2407.10606*, 2024.
- [6] J. Wang, W. Ouyang, S. Fang, Y. Zhang, X. Wu, and Z. Yi. Temptrans-mil: A general approach to enhancing multimodal tactile-driven robotic manipulation classification tasks. *IEEE/ASME Transactions on Mechatronics*, 2025.
- [7] A. George, S. Gano, P. Katragadda, and A. B. Farimani. Vital pretraining: Visuo-tactile pre-training for tactile and non-tactile manipulation policies. *arXiv preprint arXiv:2403.11898*, 2024.
- [8] A. Billard and D. Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019.
- [9] J. Cui and J. Trinkle. Toward next-generation learned robot manipulation. *Science robotics*, 6(54):eabd9461, 2021.
- [10] 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*, 3(4):3300–3307, 2018.
- [11] S. Cui, R. Wang, J. Wei, J. Hu, and S. Wang. Self-attention based visual-tactile fusion learning for predicting grasp outcomes. *IEEE Robotics and Automation Letters*, 5(4):5827–5834, 2020.
- [12] R. Feng, D. Hu, W. Ma, and X. Li. Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation. *arXiv preprint arXiv:2408.01366*, 2024.
- [13] X. Xiao, J. Liu, Z. Wang, Y. Zhou, Y. Qi, S. Jiang, B. He, and Q. Cheng. Robot learning in the era of foundation models: A survey. *Neurocomputing*, page 129963, 2025.

- [14] Z. Xu, K. Wu, J. Wen, J. Li, N. Liu, Z. Che, and J. Tang. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385*, 2024.
- [15] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- [16] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- [17] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [18] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.
- [19] Y. Tang, S. Zhang, X. Hao, P. Wang, J. Wu, Z. Wang, and S. Zhang. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778*, 2025.
- [20] D. Li, Y. Jin, Y. Sun, H. Yu, J. Shi, X. Hao, P. Hao, H. Liu, F. Sun, J. Zhang, et al. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*, 2024.
- [21] P. Ding, J. Ma, X. Tong, B. Zou, X. Luo, Y. Fan, T. Wang, H. Lu, P. Mo, J. Liu, et al. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.
- [22] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025.
- [23] H. Tan, Y. Ji, X. Hao, M. Lin, P. Wang, Z. Wang, and S. Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.
- [24] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.
- [25] X. Ma, S. Patidar, I. Haughton, and S. James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18081–18090, 2024.
- [26] S. Li, R. Krohn, T. Chen, A. Ajay, P. Agrawal, and G. Chalvatzaki. Learning multimodal behaviors from scratch with diffusion policy gradient. *Advances in Neural Information Processing Systems*, 37:38456–38479, 2024.
- [27] K. F. Gbagbe, M. A. Cabrera, A. Alabbas, O. Alyunes, A. Lykov, and D. Tsetserukou. Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2864–2869. IEEE, 2024.
- [28] M. Zhu, Y. Zhu, J. Li, Z. Zhou, J. Wen, X. Liu, C. Shen, Y. Peng, and F. Feng. Objectvla: End-to-end open-world object manipulation without demonstration. *arXiv preprint arXiv:2502.19250*, 2025.

- [29] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang, L. Guo, W. Meng, X. Zhang, et al. Multimodal fusion and vision-language models: A survey for robot vision. *arXiv preprint arXiv:2504.02477*, 2025.
- [30] J. Tu, H. Fu, F. Yang, H. Zhao, C. Zhang, and H. Qian. Texttoucher: Fine-grained text-to-touch generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7455–7463, 2025.
- [31] N. Cheng, C. Guan, J. Gao, W. Wang, Y. Li, F. Meng, J. Zhou, B. Fang, J. Xu, and W. Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *arXiv preprint arXiv:2406.03813*, 2024.
- [32] R. Feng, J. Hu, W. Xia, T. Gao, A. Shen, Y. Sun, B. Fang, and D. Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025.
- [33] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024.
- [34] S. Yu, K. Lin, A. Xiao, J. Duan, and H. Soh. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*, 2024.
- [35] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *arXiv preprint arXiv:2501.04693*, 2025.
- [36] P. Hao, C. Zhang, D. Li, X. Cao, X. Hao, S. Cui, and S. Wang. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*, 2025.
- [37] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- [38] H. Liu, Y. Yu, F. Sun, and J. Gu. Visual–tactile fusion for object recognition. *IEEE Transactions on Automation Science and Engineering*, 14(2):996–1008, 2016.
- [39] S. Cui, R. Wang, J. Wei, F. Li, and S. Wang. Grasp state assessment of deformable objects using visual-tactile fusion perception. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 538–544. IEEE, 2020.
- [40] J. Li, S. Dong, and E. Adelson. Slip detection with combined tactile and visual information. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7772–7777. IEEE, 2018.
- [41] F. Liu, C. Li, Y. Qin, A. Shaw, J. Xu, P. Abbeel, and R. Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.
- [42] 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. *arXiv preprint arXiv:2503.02881*, 2025.
- [43] L. Zhang, X. Hao, Q. Xu, Q. Zhang, X. Zhang, P. Wang, J. Zhang, Z. Wang, S. Zhang, and R. Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451*, 2025.
- [44] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. *arXiv preprint arXiv:2410.24091*, 2024.

- [45] Y. Wu, H. Lyu, Y. Tang, L. Zhang, Z. Zhang, W. Zhou, and S. Hao. Evaluating gpt-4o’s embodied intelligence: A comprehensive empirical study. *TechRxiv preprint techrxiv.174495686.69962588/v1*, 2025.
- [46] Y. Hong, Z. Zheng, P. Chen, Y. Wang, J. Li, and C. Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26406–26416, 2024.
- [47] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.
- [48] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.
- [49] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.
- [50] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.
- [51] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al.  $\pi_0$ : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [52] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambeta, R. Calandra, and K. Goldberg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024.
- [53] A. Peysakhovich and A. Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.
- [54] H. Fei, S. Wu, M. Zhang, M. Zhang, T.-S. Chua, and S. Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [55] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [56] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [57] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.
- [58] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [59] C. Zhang, S. Cui, S. Wang, J. Hu, Y. Cai, R. Wang, and Y. Wang. Gelstereo 2.0: An improved gelstereo sensor with multimedium refractive stereo calibration. *IEEE Transactions on Industrial Electronics*, 71(7):7452–7462, 2023.
- [60] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

## Appendix

This supplementary material provides additional details on the proposed method and experimental results that could not be included in the main manuscript due to page limitations. Specifically, this appendix is organized as follows.

- Sec. A presents additional details of domain randomization on simulated dataset.
- Sec. B presents details of real-world robot setup for insertion task.
- Sec. C presents a comparison between VTLA and VLA model under poor lighting condition.

### A Domain Randomization on Simulated Dataset

To enhance zero-shot Sim2Real transfer performance of VTLA model, domain randomization techniques are employed during data generation to vary parameters in the simulation environment, task configurations, and both visual and tactile observations. Tab. 8 shows the randomization parameters in details.

Table 8: The settings of domain randomization parameters.

	Parameter names	Distribution
Physical	Young’s modulus (Pa)	U(1.0e5, 5.0e5)
	Poisson ratio	U(0.3, 0.48)
	Friction coefficient	U(0.2, 0.7)
Task-related	Peg offset x in gripper (mm)	U(-1.0, 1.0)
	Peg offset z in gripper (mm)	U(-1.0, 1.0)
	Contact depth (mm)	U(0.6, 0.9)
Tactile images	Color jittering (adjustment of brightness, contrast, saturation, and hue)	/
Vision images	Direction of environmental light source	Any direction in 3D space
	Intensity of environmental light source	U(0.2, 0.6)
	Image scale transformation	U(0.9, 1.1)
	Image translation transformation (pixels)	U(-10, 10)
	Image rotation transformation (deg)	U(-3, 3)
	Image shearing transformation (deg)	U(-3, 3)
	Color jittering, Gaussian noise, motion blur, etc.	/

U(low, high) indicates a uniform distribution.

- **Physical parameters:** Young’s modulus, Poisson ratio, and the friction coefficient are randomized. This is motivated by the fact that aging of the silicone layer in the visuotactile sensor can alter these parameters over time, and accurately determining their values in the real world is highly challenging. Furthermore, the friction coefficient in physical environments is influenced by numerous factors, such as material properties, ambient humidity, and so on, making precise modeling in simulation inherently difficult.
- **Task-related:** In real-world scenarios, the robot’s grasping position and applied force on the peg may vary across episodes. To improve the policy’s adaptability, we randomize the in-hand position of the peg and the gripper width during task initialization in the simulation environment.
- **Tactile images:** In order to reduce the sim-real gap of tactile images, color jittering is applied to the simulated tactile images, including adjustments of brightness, contrast, saturation, and hue.



- **Vision images:** Domain randomization is applied to both environmental lighting and vision images. At the start of each task, three light sources from different directions are randomly configured in the environment, with their intensities individually randomized. To reduce the sim-real gap caused by discrepancies between the wrist camera poses in the real and simulated environments, the vision images are randomly scaled, translated, rotated, and sheared. Additionally, color jittering, Gaussian noise, and motion blur are applied to further improve the generalization of VTLA model.

## B Real-World Robot Setup for Insertion Task

We use a 6-DoF UR3 robot arm with a Robotiq 2F-85 gripper for real-world insertion experiments, as shown in Fig. 5. An Intel RealSense D405 camera is mounted on the wrist to capture images, while two GelStereos 2.0 sensors [59] on the gripper’s fingertips obtain tactile observations during peg-hole collisions. The robot begins by grasping the peg from a peg holder and approaching the target hole with a randomized pose. The misalignment between the peg and hole is sampled from a range of  $-2.5$  to  $2.5$  mm along the  $x$ -axis and  $y$ -axis, and from  $-5^\circ$  to  $5^\circ$  in rotation around the  $z$ -axis. The gripper then moves down to attempt insertion. The vision image is taken at the moment when the peg and hole contact. The tactile images sequences are recorded during peg-hole collisions, with each tactile sensor capturing a sequence of 4 frames. The tactile sensing program is running on ROS at a rate of 20 PFS. Subsequently, the visual-tactile observations and a language instruction are fed into the VTLA model to generate a robot action. This process is repeated iteratively until the task is either successfully completed or deemed a failure.

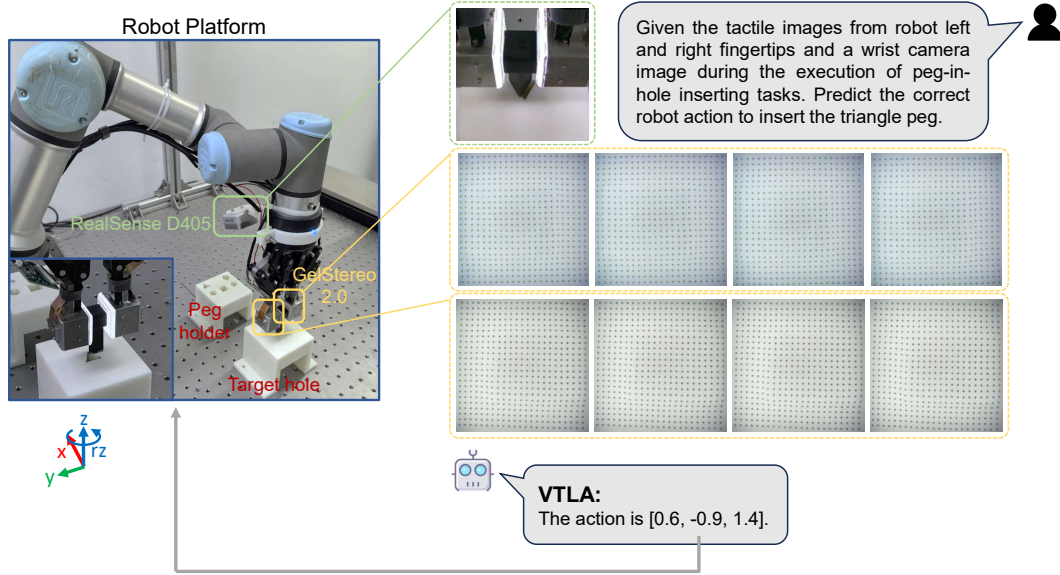


Figure 5: The insertion task setup in the real world. The left part shows the robot platform, and the right part shows the real visual-tactile observations and a dialogue round.

## C VTLA vs. VLA under Poor Lighting Condition

We compare the performance of the VTLA and VLA models on the peg-in-hole assembly task under dim lighting conditions. The procedure of peg insertion is illustrated in Fig. 6 and Fig. 7, respectively. Under dim lighting conditions, the quality of the vision image is significantly degraded (comparing the vision images in Fig. 5 and Fig. 6), making it difficult to recognize the hole position in the visual modality (steps 3–14 in Fig. 7). The VTLA model can successfully perform the insertion task under poor lighting condition, whereas VLA model struggles to complete the task. This result demonstrates that the VTLA model is effective in leveraging both visual and tactile observations, and improves the success rate of contact-rich manipulation tasks.

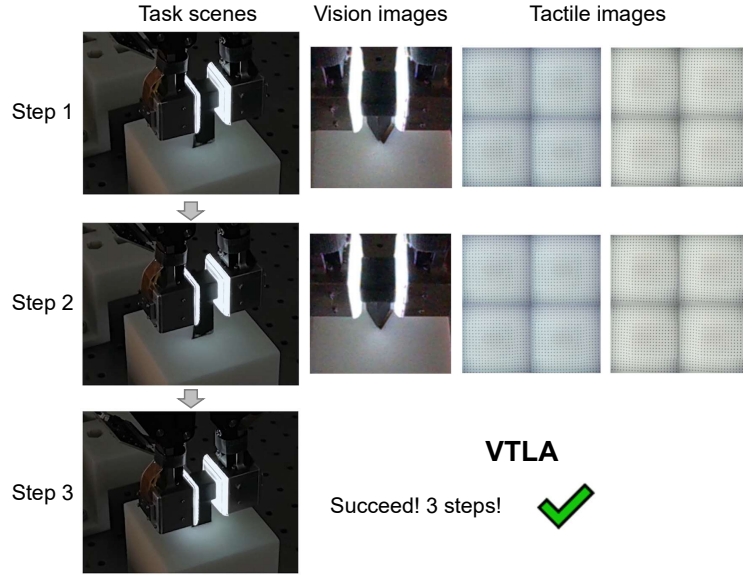


Figure 6: Snapshots of real-world insertion using the proposed **VTLA model** under poor lighting condition. The assembly clearance is 0.6 mm.

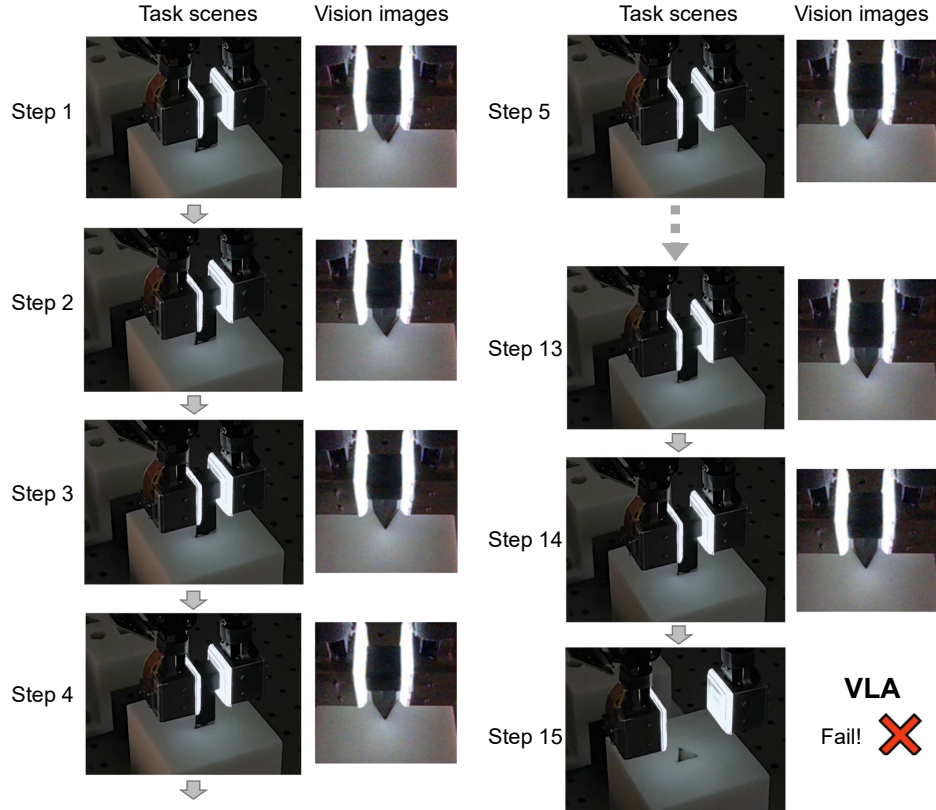


Figure 7: Snapshots of real-world insertion using the **VLA model** under poor lighting condition. The assembly clearance is 0.6 mm. The initial peg-hole misalignment is consistent with that in Fig. 6.