TLA: Tactile-Language-Action Model for Contact-Rich Manipulation

Peng Hao[1*], Chaofan Zhang[2*], Dingzhe Li[1], Xiaoge Cao[2], Xiaoshuai Hao[3], Shaowei Cui[2], and Shuo Wang[2]

Abstract— Significant progress has been made in vision-language models. However, language-conditioned robotic manipulation for contact-rich tasks remains underexplored, particularly in terms of tactile sensing. To address this gap, we introduce the Tactile-Language-Action (TLA) model, which effectively processes sequential tactile feedback via cross-modal language grounding to enable robust policy generation in contact-intensive scenarios. In addition, we construct a comprehensive dataset that contains 24k pairs of tactile action instruction data, customized for fingertip peg-in-hole assembly, providing essential resources for TLA training and evaluation. Our results show that TLA significantly outperforms traditional imitation learning methods (e.g., diffusion policy) in terms of effective action generation and action accuracy, while demonstrating strong generalization capabilities by achieving over 85% success rate on previously unseen assembly clearances and peg shapes. We publicly release all data and code in the hope of advancing research in language-conditioned tactile manipulation skill learning. Project website: **https://sites.google.com/view/tactile-language-action/**.

I. INTRODUCTION

Tactile perception is crucial for contact-rich robotic manipulation tasks [1]. For example, in fine assembly tasks, robots need to precisely sense small variations in the surface of an object [2], [3]. Tactile sensing allows robots to make subtle adjustments in their contact pose, avoiding damage or misalignment [4]. This precise perception of contact is indispensable in many complex tasks [5]. Previous studies have shown that the integration of tactile feedback significantly enhances the flexibility and robustness of manipulation skills [6], [7], especially when dealing with complex or unforeseen contact environments, thereby improving the robot's adaptability. However, current methods largely rely on specialized models trained on specific datasets [8], [9], which are limited in terms of generalization and cannot match the capabilities of general-purpose models [10], [11].

Recently, large language models have made significant breakthroughs in human-like reasoning [12], with rapid progress in the development of vision-language-action
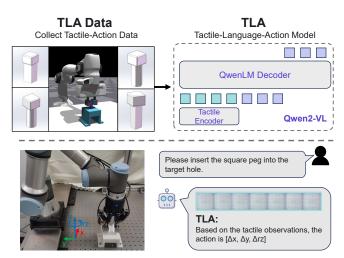
Fig. 1. Overview of the proposed Tactile-Language-Action model and tactile-action dataset. The Tactile-Language-Action model is a 7B multimodal language model trained on our tactile-action dataset collected for contact-rich manipulations. With supervised fine-tuning, the TLA can control the robot completing various peg-in-hole assembly tasks.

(VLA) models [13], [14]. By cross-modal language grounding, VLA models outperform traditional imitation learning methods, particularly in terms of generalization across different robot platforms and task settings. However, most current VLA models are primarily focused on visual tasks [15], lacking crucial tactile modalities, which limits their applicability in contact-rich manipulation tasks [16].

Despite recent efforts to align language and tactile for perceptual tasks [17], [18], [19], these studies rely on datasets that either exclude robotic actions or treat tactile as a supplementary modality [20], limiting their applicability for policy training, confined to perception-only or pick-and-place grasping tasks. The challenges in language-grounded tactile skill learning include: 1) the lack of specialized tactile-action instruction datasets for contact-rich manipulation tasks; and 2) the absence of suitable tactile-language-action models.

To address the aforementioned challenges, we construct a tactile-action instruction dataset tailored for a fingertip tactile peg-in-hole assembly scenario [21]. We propose a fine-tuning method for generalist robotic policy learning dubbed the Tactile-Language-Action (TLA) model, demonstrating that cross-modal fine-tuning enables the acquisition of generalized tactile skills through language grounding. The overview of the proposed TLA dataset and model is shown in Fig. 1.

Unlike VLA models, TLA models designed for contact-rich manipulation must effectively handle the sequential tactile information generated by a single contact action. Our key contribution is encoding the generated sequential

tactile images into a composite tactile image, which is then processed by an image encoder. This representation, combined with language-grounded reasoning, enables the model to generate robotic actions, allowing our policy to execute interactive contact and collision tasks based on natural language inference. In the fingertip-based peg-in-hole assembly task, our approach can follow natural language operation instructions, such as: "Please insert the square peg into the target hole based on the tactile observation."

Our results show that TLA outperforms traditional imitation learning methods in terms of single-step motion accuracy and achieves a significant lead in the success rate of actual shaft-hole assembly (multi-steps). Notably, in terms of generalization, TLA demonstrates strong adaptability to varying assembly clearances and peg geometries. Trained solely on assembly data with a 2.0 mm clearance, TLA achieves over 85% assembly success rates on tasks with 1.6 mm and 1.0 mm clearances. In our experiments, we collected a dataset comprising 24k tactile sequences and corresponding robot action trajectories, generated using a high-fidelity tactile simulator. To the best of our knowledge, the proposed TLA model represents the first language-grounded tactile-action generation framework, marking a significant step toward embodied intelligence.

II. RELATED WORK

A. VLM for Robot Manipulation

Due to the strong perception and reasoning abilities of VLM, RT-2 [11] treats actions in the manipulation dataset as tokens. It fine-tunes on VLM and calls the trained model the Vision-Language-Action (VLA) Model. However, RT-2 is closed-source. RoboFlamingo [22] and OpenVLA [13], which follow similar concepts to RT-2, are open-sourced in the robotics community. GR-1 [23] and GR-2 [24] leverage an Internet-scale video dataset for pre-training. They are then fine-tuned using the manipulation dataset. This method incorporates the physics and dynamics information from the Internet-scale video dataset into the VLA. RDT-1B [25] and PI [26] integrate diffusion concepts into the VLA model framework, demonstrating the effectiveness of training with a diffusion objective. However, the VLA models mentioned above are vision-based and are limited to simple push, pull, and grasp-and-place tasks. This paper explores the effectiveness of using tactile input as an observation for contact-rich peg-in-hole tasks.

B. Tactile-Language Model in Robotics

Recent studies have attempted to combine tactile information with language models [17] [19]. Fu et al. [19] build a texture-tactile dataset using ChatGPT and propose the Tactile-Vision-Language Model to explore language models' ability in material understanding. However, the dataset scale limits the model's generalization in real-world scenarios. To address this, Cheng et al. [18] propose Touch100k, a large-scale dataset for tactile-vision-based material classification. Unfortunately, these works have not applied tactile information to robot manipulation. Jones et al. [20] propose FuSe,

integrating tactile information into the VLA model to acquire a multimodal robot policy. Different from these works, our work focuses on establishing a connection between tactile perception and actions based on a pre-trained language model. To the best of our knowledge, TLA is the first language-action model solely based on tactile perception.

III. DATASET

In this section, we introduce a tactile-action instruction dataset collected in the process of peg-in-hole assembly. In this task, the robot equipped with GelStereo 2.0 visuotactile sensors [27] attempts to insert a peg into the corresponding hole based on fingertip tactile sensing and language instruction. For efficient data collection, we build a simulation environment for this task in NVIDIA Isaac Gym. The deformation of the visuotactile sensor during interaction is simulated based on Finite Element Method (FEM) using a Flex physics engine. The tactile imprint rendering method proposed in [28] is utilized to simulate tactile images. To narrow the sim-real gap, a tactile image obtained from a real sensor is employed for texture mapping instead of a manually designed pattern. In this way, we can obtain high-fidelity tactile images during insertion attempts.

The process of the peg-in-hole task is described as follows. The gripper first grasps a peg with a shape description and moves to the top of the corresponding hole with a random 3-DOFs misalignment in the x-axis, y-axis, and rotation around the z-axis (denoted by rz). Then, the gripper moves down to attempt insertion. If a collision occurs between the peg and hole during the downward movement, this attempt is deemed failed, and the gripper lifts up waiting for the next attempt. The tactile image sequence during collision is recorded to infer the robot action $(\Delta x, \Delta y, \Delta rz)$ that adjusts the peg pose. If the collision does not occurred while the gripper moving down to a predetermined position, the task is deemed successful. The maximum number of attempts is 15. Otherwise, the task fails.

In simulation, we employ a random insertion policy for this peg-in-hole task. For each attempt, we save the tactile image sequence and the peg-hole pose error. Then, we create the action labels $(\Delta \hat{x}, \Delta \hat{y}, \Delta \hat{rz})$ from peg-hole pose errors (e_x, e_y, e_{rz}).

$$\Delta \hat{x} = \begin{cases} \mathbb{F}_{clip}(-e_x + c/2, -\delta, 0), & if \ e_x \geq 0, \\ \mathbb{F}_{clip}(-e_x - c/2, 0, \delta), & if \ e_x > 0, \end{cases} \quad (1)$$

$$\Delta \hat{y} = \begin{cases} \mathbb{F}_{clip}(-e_y + c/2, -\delta, 0), & if \ e_y \geq 0, \\ \mathbb{F}_{clip}(-e_y - c/2, 0, \delta), & if \ e_y > 0, \end{cases} \quad (2)$$

$$\Delta \hat{rz} = \mathbb{F}_{clip}(-e_{rz}, -1.5°, 1.5°) \quad (3)$$

where c is the assembly clearance. $\mathbb{F}_{clip}$ limits the action to a certain range to improve the stability of the policy. δ is set to 1 mm according to our experience. In this paper, the dataset is collected only using pegs with 2.0 mm clearance.

To facilitate model training, the collected interaction data is transformed into the instruction format. The <|im_start|> and <|im_end|> tokens mark the start and end of each dialogue round. Tactile images are input with
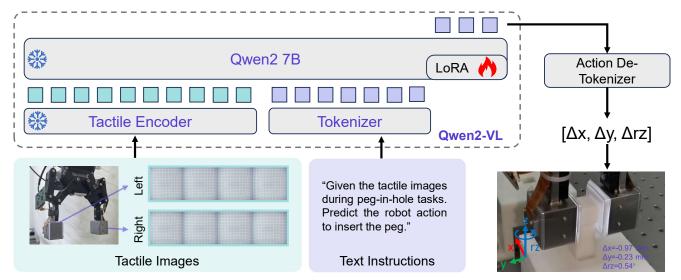
Fig. 2. Architecture of Tactile-Language-Action model. Given the tactile images and text instruction, TLA predicts the 3-dimensional robot actions. TLA consists of two modules: (1) Tactile Encoder captures the tactile representations for the language model. (2) The Language Model fuses and understands tactile and text information to predict the robot's actions.

Fig. 3. An example of TLA Dataset.

$\texttt{<|vision\_start|>}$ and $\texttt{<|vision\_end|>}$, denoting start and end of visual inputs. A text instruction defines the task, clarifies the peg type, and poses requirements. The robot's actions during data collection are saved as ground-truth. An example of the TLA data is shown in Fig. 3.

IV. TACTILE-LANGUAGE-ACTION MODEL

This section introduces the proposed Tactile-Language-Action model, as shown in Fig 2. TLA is built based on Qwen2-VL [29], which includes a Vision Transformer (ViT) [30] for encoding the visual input and the Qwen2 language model [31] for understanding multimodal information and generating text. This section first gives the details of encoding tactile information with ViT, then introduces how to predict robot actions with the language model.

A. Tactile Encoder

As mentioned in the previous section, the tactile information of visuotactile sensor is represented in the form of images. During robot manipulation, the tactile images continuously change according to the robotic gripper's contact state. Therefore, the tactile encoder needs to process two temporally aligned image sequences.

The tactile encoder's challenge lies in extracting temporal variation from input tactile images. To address this, we synthesize two tactile image sequences into a single image, converting temporal information into spatial space to facilitate ViT-based feature extraction. Specifically, the input image set denotes as $I = \{I_l^t, I_r^t; t = 0, 1, 2, 3\}$, where I_l^t and I_r^t represent the tactile images of the left and right fingertip at timestamp t, respectively. These eight images are arranged in a 3×3 grid, with the last grid filled by a white image, and are resized to 616×616 as model input.

We use the ViT of Qwen2-VL as the tactile encoder, which is shown in Fig. 2. The concatenated tactile images are processed by the tactile encoder to obtain tactile features. Moreover, a Multi-Layer Perceptron (MLP) layer is used to further compress the tactile features within a 2×2 range to a single token. Therefore, for the input tactile image $I \in \mathbb{R}^{616 \times 616}$, 1936 tactile tokens can be obtained after passing through the ViT with a patch size of 14.

B. Action Prediction with Language Model

This section gives the details of using the Qwen2 language model to predict robot actions, as shown in Fig. 2. The inputs of the language model are tactile tokens and language tokens, which are acquired by encoding the original multimodal inputs through the tactile encoder and the tokenizer, respectively. The Qwen2 7B is TLA's backbone model and is fine-tuned on the TLA dataset.

The encoding of continuous numbers with a discrete tokenizer affects the model's performance in number-sensitive tasks [32]. Previous work [13] simply overwrite the least used tokens as "special tokens" and assigns bin IDs to each token. Unlike prior approaches, we retain the numerical encoding scheme to ensure that the pre-training acquired numerical knowledge is effectively leveraged in robotic manipulation tasks.

However, the tokenizer of Qwen2 encodes numbers individually. Since there are many decimals in the robot action data, this redundant information increases the difficulty of the model training. To this end, we simplify the ground-

truth by scaling all actions with a ratio and rounding to integers. Specifically, the processing procedure is calculated as $A_{gt} = A_{raw} \cdot s$, where $A_{gt} \in \mathbb{R}^3$, $A_{raw} \in \mathbb{R}^3$, $s \in \mathbb{R}^3$ are the ground-truth action, raw action and scale factors.

C. Training and Inference

Previous work has shown that the VLMs achieve better performance with the frozen visual encoder during training [29] [33]. Therefore, we freeze the parameters of the Tactile Encoder during fine-tuning. Additionally, we use Low-Rank Adaptation (LoRA) [34] to efficiently fine-tune the Qwen2 7B language model. The ground-truth actions are used as labels to calculate the Next Token Prediction loss, which is calculated as follows,

$$\mathcal{L}_{\text{NTP}} = -\sum_{t=1}^{T} p(y_t) \log P(y_t \mid y_{<t}, x) \qquad (4)$$

where T and y_t are the length of the action sequence and the ground-truth token at step t, respectively. $y_{<t}$ denotes the previously generated tokens, replaced with ground-truth during training. x represents the input, including tactile image tokens and text instruction tokens.

During inference, TLA sequentially predicts the probability distribution of the robot's actions based on the input tactile observations and instruction texts. The generation process is carried out through beam search until the end token is generated. Finally, the Action-De-Tokenizer maps all the generated probabilities to natural language text according to the vocabulary and converts them into floating-point numbers that can be executed by the robot.

V. EXPERIMENT

In this section, we investigate the effectiveness of our proposed TLA model for language-conditioned tactile learning in a fingertip peg-in-hole assembly task. Specifically, we focus on using only tactile observations to generate assembly action commands using TLA, enabling language-conditioned tactile skill learning. The goal of our experiment is to answer the following questions:

- Compared to conventional imitation learning approaches that train from scratch, does our language-conditioned TLA model achieve a better understanding and modeling of the relationship between tactile observations and task-specific action commands?
- Can language models enhance the generalization capability of TLA across different objects and variations of the peg-in-hole assembly task?
- Is the trained TLA model capable of effectively controlling the robot to successfully perform a variety of peg-in-hole assembly tasks with different geometries and constraints?

A. Baseline and Metrics

To answer the above questions, we compare the following baseline methods and ablation methods with the proposed TLA on various tasks.

Method	GCR (%)↑	L1 x (mm)↓	L1 y (mm)↓	L1 rz (deg)↓
BC [35]	10.4	0.803	0.302	0.205
DP [37]	8.5	0.370	0.382	0.568
SP-TLA (Ours)	**12.5**	**0.079**	**0.122**	**0.173**

- **Behavior Cloning (BC)** [35]: The ResNet-50 [36] is employed as the policy network. The network takes the tactile image as input and outputs the robot action, which is trained using a supervised manner.
- **Diffusion Policy (DP)** [37]: The diffusion policy in [37] utilizes the conditional denoising diffusion process to learn the peg-in-hole assembly policy.
- **Single-Peg TLA (SP-TLA)**: The TLA model trained on the square peg insertion dataset.
- **Multi-Peg TLA (MP-TLA)**: The TLA model trained on the square and triangular peg insertion dataset.

To evaluate different policies, we first evaluate the performance of the model's effective action generation with **Goal Convergence Rate (GCR)**, which is defined as the percentage of all the output actions that are correct in the x, y, and rz directions. Then, the L1 distance is used to evaluate the performance of different models, which is calculated as follows,

$$L_1 = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|, \qquad (5)$$

where m, y_i, $\hat{y}_i$ are the sample numbers in GCR, action prediction, and ground-truth, respectively. We calculate the L1 distance of each action dimension and present the results.

B. Comparison on Single-Peg Inserting Tasks

To answer the first question, this subsection compares the TLA with the baseline methods on the TLA dataset.

Experimental Setup We take 8k square peg insertion data from the TLA dataset to compare the performance of TLA and the baseline methods. Specifically, the training set and the test set are split into 6k and 2k, respectively. Each training sample consists of 8 tactile images, which correspond to the left and right fingertips with a time length of 4. The label is the robot action $(\Delta x, \Delta y, \Delta rz)$, which represents the movement in the horizontal plane and the rotation around the z-axis, respectively. The TLA model is trained on 8 Nvidia A6000 GPUs for 20 epochs.

Results The performance of different methods on the single-peg test set is shown in Table I. The GCR results indicate that TLA has more correct actions than the baselines. Furthermore, the L1 results show that the correct actions predicted by TLA have more accurate step lengths than others, where the L1 of the x-direction is reduced by 78% compared to the second-best method. The superior L1 performance implies that TLA is expected to complete tasks with fewer operation steps, demonstrating better manipulation efficiency.

The visualization results of model predictions are shown in Fig. 4, which includes three sub-figures by combining different action dimensions. The starting point of each arrow
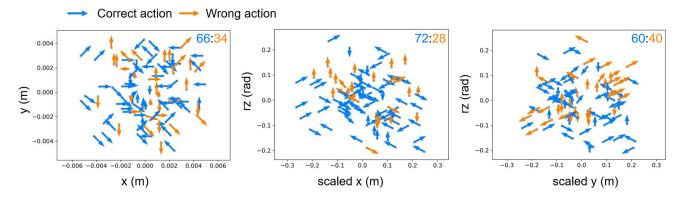
Fig. 4. Visualization of action prediction on the single-peg test set. By combining each dimension's actions in pairs, three sub-figures are given, where the correct and wrong actions are marked with different colors. Scaled means the raw data is scaled by a factor of 50 for visualization. The results show that TLA can predict the correct actions for various deviations.

TABLE II

COMPARISON OF DIFFERENT METHODS ON THE MULTI-PEG TEST SET.

Method	ID				OOD			
	GCR(%)↑	L1 x (mm)↓	L1 y (mm)↓	L1 rz (deg)↓	GCR (%)↑	L1 x (mm)↓	L1 y (mm)↓	L1 rz (deg)↓
BC [35]	**18.4**	0.260	0.655	0.186	0.152	0.286	0.722	0.246
DP [37]	7.4	0.371	0.348	0.480	0.080	0.386	0.369	0.544
MP-TLA (Ours)	**18.4**	**0.102**	**0.114**	**0.135**	**0.165**	**0.121**	**0.102**	**0.184**

represents the pose deviation between the current peg and the correct insertion position, while the arrow direction indicates the predicted action movement. The correct and wrong predictions are marked in blue and orange, respectively. For the x-y pair, an action is considered correct if it reduces the position deviation. For the pairs of x-rz and y-rz, an action is correct if it reduces the deviation in at least one dimension. Experimental results demonstrate that TLA's actions perform well in planar translation, effectively guiding the robot toward the target hole. However, in the translation-rotation visualization, more incorrect actions are observed along the y-axis. We attribute this to the limitation of 2D tactile images, which effectively represent tactile information parallel to the gripper (x-axis) but poorly represent information perpendicular to the gripper (y-axis). This imbalanced feature representation increases the difficulty of y-axis predictions.

C. Comparison on Multi-Peg Inserting Tasks

To address the second question, this subsection trains TLA on various peg types and compares it with baseline methods.

Experimental Setup We take 16k samples of square and triangular pegs for training, equally split. An additional 8k samples are collected for evaluation: 4k for square/triangular pegs and 4k for round/hexagonal pegs. The input, output, and metrics align with the single-peg task. The training uses 8 Nvidia A6000 GPUs for 10 epochs.

Results The comparison of different methods on the multi-peg test set is presented in Table II. In the in-distribution (ID) experiments, TLA achieves the lowest L1 error on seen pegs, indicating that the actions predicted by TLA are not only correct but also exhibit more precise and stable step lengths during execution. For the out-of-distribution (OOD) setting, TLA maintains performance that is closely aligned

TABLE III

COMPARISON OF DIFFERENT METHODS IN INSERTION TASKS WITH VARIOUS CLEARANCES.

Method	2.0 mm		1.6 mm		1.0 mm		Total	
	Suc↑	Step↓	Suc↑	Step↓	Suc↑	Step↓	Suc↑	Step↓
BC [35]	44	2.60	32	4.00	18	4.44	31	3.68
DP [37]	58	**2.45**	43	4.35	20	4.70	40	3.83
SP-TLA	**96**	3.15	**94**	3.77	74	4.37	88	3.76
MP-TLA	94	3.04	86	**3.30**	**90**	4.35	**90**	**3.56**

with that on the ID set, highlighting its strong generalization capabilities when facing unseen peg geometries. Notably, TLA-OOD does not suffer from any significant degradation compared to TLA-ID, in stark contrast to traditional imitation learning baselines, which exhibit pronounced performance drops under distribution shifts. This demonstrates that TLA is remarkably robust to variations in peg shapes and is capable of effectively transferring its learned manipulation skills to novel, unseen configurations, substantially outperforming conventional approaches in terms of generalization.

D. Robotic Insertion Tasks

In this subsection, the manipulation performance of different models is evaluated in the robotic insertion tasks to answer the third question.

Experimental Setup We test the manipulation performance of different models in two types of peg-in-hole assembly tasks. First, we evaluate the model performance with different assembly clearances, including the square peg-in-hole assembly tasks with assembly clearances of 2.0 mm, 1.6 mm, and 1.0 mm, respectively. Then, we report the performance of different models on square, triangular, round, and hexagonal peg insertion tasks, where all assembly clearances are set to 2.0 mm. Each task is repeated 50 times to calculate the final results.
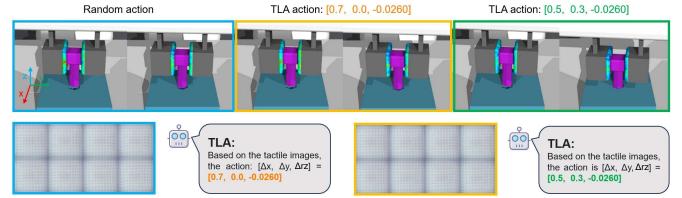
Fig. 5. Snapshots of the robot successfully inserting the hexagonal peg into the target hole with the assistance of our proposed TLA.

TABLE IV

COMPARISON OF DIFFERENT METHODS IN INSERTION TASKS WITH

VARIOUS PEG TYPES.

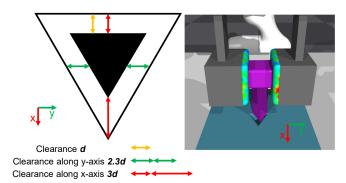| Method | ID | | | | OOD | | | |
| | Square | | Triangle | | Round | | Hexagon | |
	Suc↑	Step↓	Suc↑	Step↓	Suc↑	Step↓	Suc↑	Step↓
BC [35]	60	3.37	84	2.67	62	2.84	66	**2.30**
DP [37]	60	3.60	60	3.40	56	2.89	54	2.37
SP-TLA	**96**	3.15	76	2.82	92	2.80	90	2.78
MP-TLA	94	**3.04**	**92**	**1.80**	**94**	**2.66**	**96**	3.00

Fig. 6. Visualization and analysis of failure case. The TLA fails to insert the triangular peg because the triangular peg has a smaller assembly clearance along the y-axis, and the TLA model based on 2D observations has poor prediction performance along the y-axis.

At the beginning of each episode, the robot has already grasped the peg in its gripper, and the end-effector is randomly set at a starting position near the target hole. Then, the robot attempts the first insertion and obtains tactile images. Subsequently, TLA or other models predict the robot's action based on the tactile observation and control the robot to insert again. The robot continues attempting until the insertion is successful or reaches the maximum attempt number 15. The success rate and the average steps of successful episodes are used to evaluate the performance of all models.

Results The quantitative comparison results of different methods on various assembly clearances are shown in Table III. Compared with the baseline methods, TLA has achieved a superior success rate and manipulation efficiency, surpassing the second-best method by 50% in the success rate. Even with the challenging assembly clearance of 1.0 mm, our TLA model also achieves superior performance than baselines, showing its excellent generalization for assembly clearances. The quantitative comparison results of different methods on various peg types are shown in Table IV. Both the two TLA models have achieved a better success rate and steps than baseline methods, demonstrating excellent generalization performance across different peg types. In particular, the experiments on OOD set show that the MP-TLA model performs better than the SP-TLA model. We attribute this to the training data of different pegs can further enhance the generalization of TLA.

Visualization The visualization result of TLA controlling the robot to successfully insert the peg is shown in Fig. 5. First, the robot fails to insert the peg with the initial action. Then, TLA predicts the adjusted robot actions based on the

tactile observation during previous attempts. Finally, TLA successfully controls the robot to approach and insert the peg after two rounds of attempts.

The visualization of a failure case is shown in Fig. 6, where TLA fails to control the robot to insert the triangular peg into the hole. We attribute this failure case to the triangular hole having different allowable deviations along the x-axis and the y-axis. Specifically, when assembly clearance is d, the allowable assembly deviation along the x-axis is 3d, while that in the y-axis is only 2.3d. Moreover, since the current tactile information is 2D images, its representation in the direction perpendicular to the gripper (i.e. y-axis) is relatively poor than the other two axes. Therefore, this imbalanced feature representation increases the difficulty for the model to understand the contact information along the y-axis. As a result, due to the model's poor performance in the y-axis direction and the smaller allowable deviation of the triangular shape in the y-axis direction, TLA exhibits poor performance in triangular peg insertion tasks.

VI. DISCUSSION AND LIMITATIONS

In this study, we introduce TLA, a Tactile-Language-Action model designed for contact-rich manipulation scenarios. Through a cross-modal fine-tuning process, TLA enables the acquisition of generalized tactile skills via language grounding. We further demonstrate that TLA significantly

outperforms traditional imitation learning methods in a challenging fingertip tactile peg-in-hole assembly task, achieving superior assembly success rates. Moreover, it exhibits strong generalization capabilities across variations in assembly clearness and peg shapes.

Despite these promising results, TLA still has several limitations. One notable limitation is that the model does not rigorously capture tactile temporal information. Instead, it encodes temporal cues through spatial arrangements, which may not fully exploit the sequential nature of tactile data. Future work should explore more effective representations and encoding strategies tailored to sequential tactile perception. Another area for improvement lies in the selection of tactile signal formats, which remains relatively rudimentary in this study. Different tactile representations, such as 2D tactile images, 2D contact depth maps, and 3D tactile point clouds, offer unique advantages, and investigating their integration with VLMs for tactile skill learning could further enhance the model's capabilities. Additionally, the current action detokenization process is relatively simplistic, leaving room for refinement. A more sophisticated decoding mechanism could improve the interpretability and precision of learned actions, making it a valuable direction for joint exploration between VLA and TLA models.

We will deploy TLA in real-world environments to evaluate its sim-to-real generalization, testing how well the policy learned in simulation transfers under real-world uncertainties. All related data, models, and code are available on our project website, where we will also share updates on our real-world experiments. We welcome the community to follow our progress and explore these challenges together.

REFERENCES

[1] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[2] S. Cui, R. Wang, J. Hu *et al.*, "In-hand object localization using a novel high-resolution visuotactile sensor," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6015–6025, 2021.

[3] J. Cui and J. Trinkle, "Toward next-generation learned robot manipulation," *Science robotics*, vol. 6, no. 54, p. eabd9461, 2021.

[4] R. Calandra, A. Owens, D. Jayaraman *et al.*, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[5] Q. Li, O. Kroemer, Z. Su *et al.*, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[6] C. Wang, S. Wang, B. Romero *et al.*, "Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," in *International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5633–5640.

[7] H. Qi, B. Yi, S. Suresh *et al.*, "General in-hand object rotation with vision and touch," in *Conference on Robot Learning*, 2023, pp. 2549–2564.

[8] M. A. Lee, Y. Zhu, P. Zachares *et al.*, "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.

[9] S. Dong, D. K. Jha, D. Romeres *et al.*, "Tactile-rl for insertion: Generalization to objects of unknown geometry," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6437–6443.

[10] O. M. Team, D. Ghosh, H. Walke *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[11] A. Brohan, N. Brown, J. Carbajal *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[12] H. Naveed, A. U. Khan, S. Qiu *et al.*, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[13] M. J. Kim, K. Pertsch, S. Karamcheti *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[14] J. Wen, Y. Zhu, J. Li *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," *arXiv preprint arXiv:2409.12514*, 2024.

[15] A. O'Neill, A. Rehman, A. Maddukuri *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6892–6903.

[16] H. Zhou, X. Yao, Y. Meng *et al.*, "Language-conditioned learning for robotic manipulation: A survey," *arXiv preprint arXiv:2312.10807*, 2023.

[17] F. Yang, C. Feng, Z. Chen *et al.*, "Binding touch to everything: Learning unified multimodal tactile representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 340–26 353.

[18] N. Cheng, C. Guan, J. Gao *et al.*, "Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation," *arXiv preprint arXiv:2406.03813*, 2024.

[19] L. Fu, G. Datta, H. Huang *et al.*, "A touch, vision, and language dataset for multimodal alignment," *arXiv preprint arXiv:2402.13232*, 2024.

[20] J. Jones, O. Mees, C. Sferrazza *et al.*, "Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding," *arXiv preprint arXiv:2501.04693*, 2025.

[21] W. Chen, J. Xu, F. Xiang *et al.*, "General-purpose sim2real protocol for learning contact-rich manipulation with marker-based visuotactile sensors," *IEEE Transactions on Robotics*, pp. 1509–1526, 2024.

[22] X. Li, M. Liu, H. Zhang *et al.*, "Vision-language foundation models as effective robot imitators," *arXiv preprint arXiv:2311.01378*, 2023.

[23] H. Wu, Y. Jing, C. Cheang *et al.*, "Unleashing large-scale video generative pre-training for visual robot manipulation," *arXiv preprint arXiv:2312.13139*, 2023.

[24] C.-L. Cheang, G. Chen, Y. Jing *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[25] S. Liu, L. Wu, B. Li *et al.*, "Rdt-1b: a diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[26] K. Black, N. Brown, D. Driess *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[27] C. Zhang, S. Cui, S. Wang *et al.*, "Gelstereo 2.0: An improved gelstereo sensor with multimedium refractive stereo calibration," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 7, pp. 7452–7462, 2023.

[28] S. Cui, Y. Wang, S. Wang *et al.*, "Tactile imprint simulation of gelstereo visuotactile sensors," in *IEEE International Conference on Mechatronics and Automation (ICMA)*, 2023, pp. 650–656.

[29] P. Wang, S. Bai, S. Tan *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[31] A. Yang, B. Yang, B. Hui *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[32] L. Qian, J. Li, Y. Wu *et al.*, "Momentor: Advancing video large language model with fine-grained temporal reasoning," *arXiv preprint arXiv:2402.11435*, 2024.

[33] S. Karamcheti, S. Nair, A. Balakrishna *et al.*, "Prismatic vlms: Investigating the design space of visually-conditioned language models," in *Forty-first International Conference on Machine Learning*, 2024.

[34] E. J. Hu, Y. Shen, P. Wallis *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR*, p. 3, 2022.

[35] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," *arXiv preprint arXiv:1805.01954*, 2018.

[36] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] C. Chi, Z. Xu, S. Feng *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.