ViTaMIn: Learning Contact-Rich Tasks Through Robot-Free Visuo-Tactile Manipulation Interface

Preview

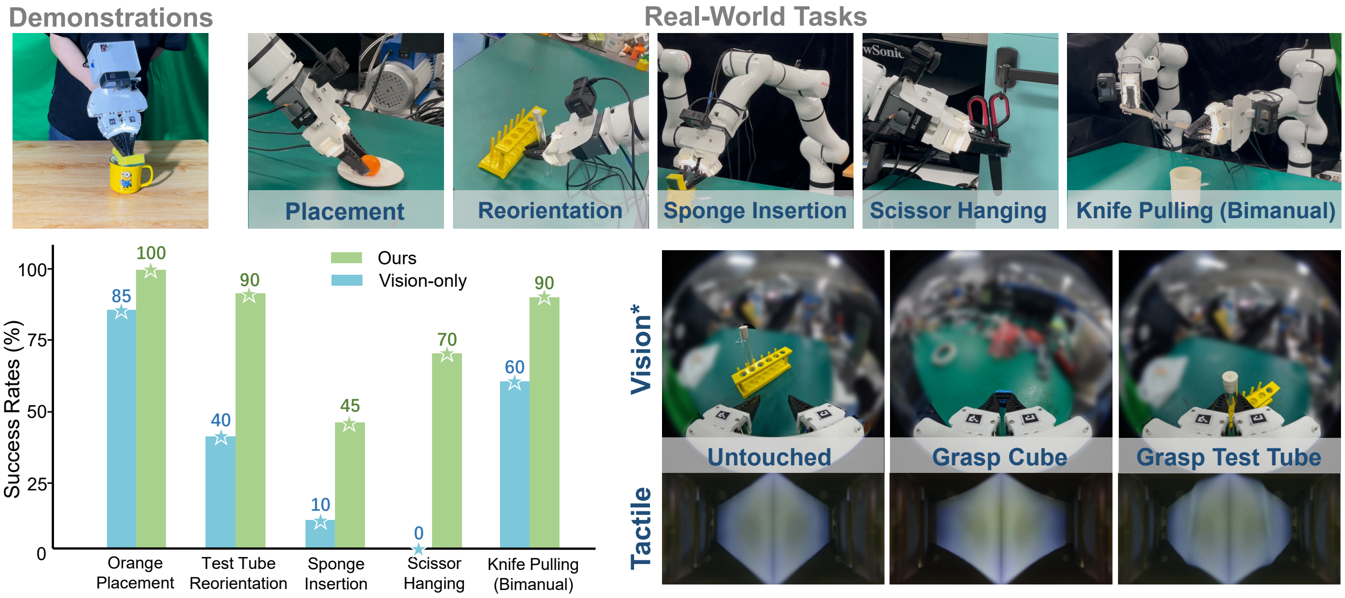

Fig. 1: ViTaMIn overview. Our system comprises a portable data collection device that integrates visual and tactile sensing, a multimodal representation learning framework for fusing visual and tactile information, and demonstrations of various contact-rich manipulation tasks. This system facilitates efficient collection of manipulation data without requiring complex robot setups. (*Backgrounds in the images are blurred.)

***Abstract*—Tactile information plays a crucial role for humans and robots to interact effectively with their environment, particularly for tasks requiring the understanding of contact properties. Solving such dexterous manipulation tasks often relies on imitation learning from demonstration datasets, which are typically collected via teleoperation systems and often demand substantial time and effort. To address these challenges, we present ViTaMIn, an embodiment-free manipulation interface that seamlessly integrates visual and tactile sensing into a hand-held gripper, enabling data collection without the need for teleoperation. Our design employs a compliant Fin Ray gripper with tactile sensing, allowing operators to perceive force feedback during manipulation for more intuitive operation. Additionally, we propose a multimodal representation learning strategy to obtain pre-trained tactile representations, improving data efficiency and policy robustness. Experiments on five contact-rich manipulation tasks demonstrate that ViTaMIn significantly outperforms baseline methods, demonstrating its effectiveness for complex manipulation tasks.**

I. Introduction

Humans rely on both visual and tactile modalities to perform a diverse range of manipulation tasks in daily life. For instance, when inserting a plug into a socket or tightening a screw, vision helps with identifying and aligning components, while tactile signals enable precise force control during contact. This seamless integration of vision and touch enhances human dexterity, particularly in tasks that require contact-rich control, handling visual occlusions, or performing in-hand manipulations.

Recent progress in learning from demonstrations [18, 2, 3, 5] has shown significant potential for advancing general-purpose robots, enabling them to efficiently acquire complex skills from human demonstrations. Consequently, developing systems to collect high-quality demonstration data has been a recent key focus. Prior work have explored real-world data collection methods, including joint-mapped devices and exoskeletons [1, 9, 42, 8], and vision-based teleoperation frameworks [4, 27]. Nevertheless, these techniques require real-time teleoperation of a physical robot during data collection, which constrains efficiency and flexibility. In contrast, portable devices [32, 7, 34, 6] present a more scalable and cost-effective alternative to collect demonstration without teleoperation. Moreover, they can be seamlessly integrated into various embodiments, providing a more flexible data collection approach. However, these portable devices primarily focus on capturing vision-only demonstration data, limiting their usage for contact-rich and dexterous manipulation tasks where tactile feedback plays a crucial role.

In this work, we aim to address both the challenge of efficient data collection and the need for learning more dexterous tasks using visuo-tactile demonstrations. To this end, we introduce ViTaMIn, a novel and effective visuo-tactile manipulation interface designed to capture high-quality demonstrations with enhanced efficiency and flexibility. Unlike conventional approaches that rely on expensive or rigid tactile sensors, ViTaMIn leverages an omnidirectional compliant Fin Ray gripper with customized tactile sensing, which can detect contact from all directions as an expressive tactile signal for robot manipulation. We integrate the tactile-aware Fin Ray gripper with UMI [6], enhancing the collected data with rich multimodal information and improving policy learning performance while maintaining the core advantages of portable devices. Additionally, our system enables operators to perceive force feedback during manipulation, facilitating more intuitive and seamless operation.

Pre-trained visual representations have shown improved performance in robotic manipulation [24, 20, 38, 29, 21], benefiting from large-scale visual pre-training. To fully leverage the visuo-tactile datasets collected with ViTaMIn, we adopt a multimodal representation learning strategy to pre-train tactile representations, enhancing the robustness and generalizability of our sensor-based policies. Our pre-training objective integrates masked autoencoding [13] and contrastive learning for multimodal alignment [28], where future image observations are aligned with masked current images and tactile signals. Through extensive experiments on five challenging contact-rich manipulation tasks, our visuo-tactile policy, enhanced by multimodal pre-training, exhibits superior data and training efficiency while demonstrating strong generalization across diverse objects and environmental conditions.

In conclusion, our contributions are:

- ViTaMIn provides a portable, scalable, and efficient visuo-tactile data collection platform in a robot-free setup. ViTaMIn achieves superior performance over vision-only baselines across five manipulation tasks by leveraging visuo-tactile demonstrations.
- ViTaMIn proposes an effective multimodal representation learning strategy, which significantly improves the data efficiency, robustness and generalization capabilities.

II. Related Work

A. Visuo-Tactile Manipulation

The integration of visual and tactile sensing is essential for robotic manipulation as it provides complementary signals about scene observations and physical contact. Early works [14, 25, 22] use RGB cameras and force/torque sensors to infer contact status for making decisions. However, the information from force/torque sensors is low-dimensional and insufficient for more dexterous manipulation tasks. More recently, vision-based tactile sensors have gained attention for their ability to capture high-resolution contact information [26, 19, 12], but the rigid design of these sensors restricts the compliance of the end effector, limiting their applicability in complex tasks. In our work, we use a Fin-Ray-shaped compliant and all-directional tactile sensor, which can detect contacts from all directions, essential for safe and robust manipulation.

The work most related to ours is Huang et al. [15], which attached flexible resistive tactile sensors with a resolution of 16×16 onto a Fin Ray gripper and devised a 3D visuo-tactile representation to integrate these two modalities, thereby enabling more efficient learning. Our work differs from theirs in three significant aspects: (1) Our data collection device is portable and low-cost. (2) Our vision-based tactile sensor has a higher resolution (640×480), which is essential for precise manipulation. (3) We pre-train effective tactile representations to enhance the generalization capability and data efficiency of our policy.

B. Data Collection System for Robot Manipulation

Recent advancements in learning from demonstrations [18, 2, 3, 5] have demonstrated promising results in developing general-purpose robots, allowing them to acquire complex skills efficiently by leveraging human demonstrations. Therefore, efficiently collecting high-quality demonstrations has become a key research focus. While simulation platforms can theoretically generate unlimited demonstrations, their fidelity remains inadequate for objects with complex physical properties [30, 16, 39, 11], limiting their applicability to real-world tasks. On the other hand, recent research has focused on developing efficient real-world data collection systems, such as devices or exoskeletons with joint-mapping [1, 9, 42], exoskeletons [8], or vision-based systems [4, 27]. However, these approaches require teleoperating a physical robot during data collection, which limits efficiency and flexibility. In contrast, portable devices [32, 7, 34, 6] offer several advantages: they are low-cost, flexible, and do not depend on a specific physical robot. Additionally, they can be seamlessly integrated into various embodiments and provide a more user-friendly experience for data collection. Unlike prior work that relies solely on visual observations, we enhance the UMI data collection system [6] by integrating tactile sensing. This addition enriches the collected data with multimodal information, improving policy learning performance while preserving the key benefits of portable devices. Moreover, our system allows operators to perceive force feedback during manipulation, enabling more intuitive and seamless operation.

C. Multimodal Pre-training for Robotics

Pre-trained visual representations have shown improved performance and generalization in robotic manipulation [24, 20, 38, 29, 21] motivated by self-supervised learning techniques [13, 28]. Similar strategies have been employed in multimodal representation learning [33, 40, 41] by integrating visual, tactile, and proprioceptive modalities, allowing robots to perceive object properties beyond visual appearance. For example, Sferrazza et al. [33] introduced a masked multimodal autoencoding framework that jointly learns visuo-

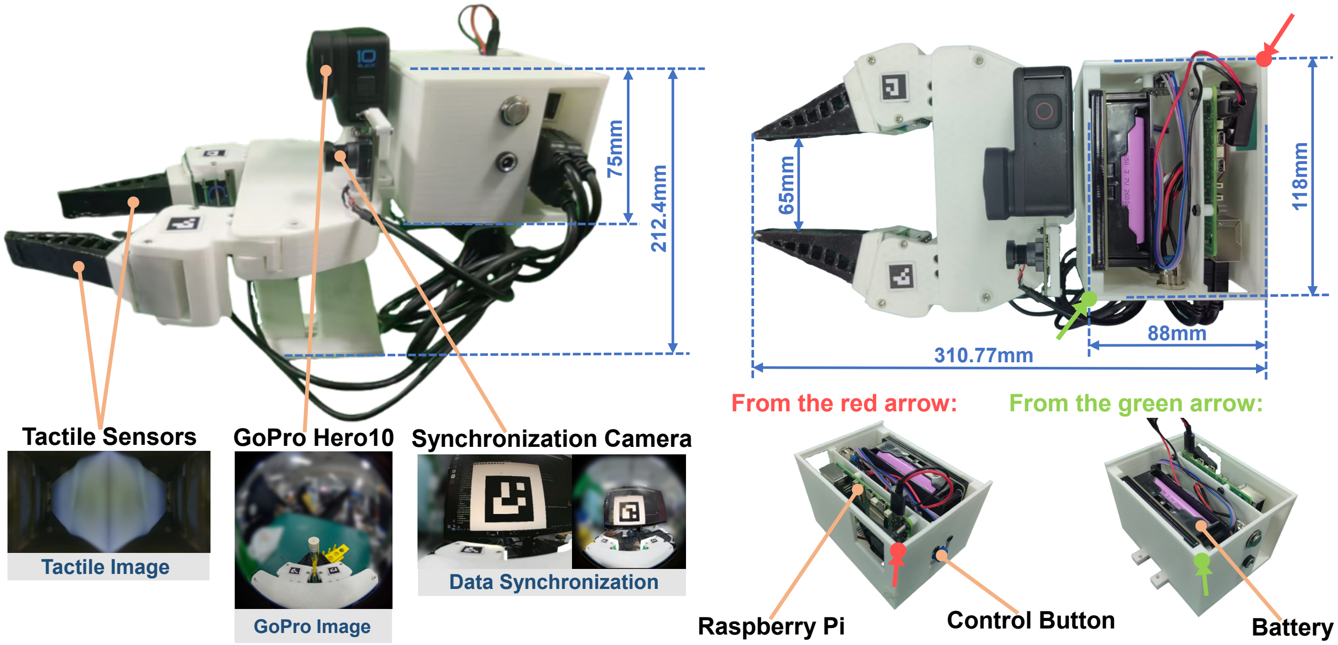

Fig. 2: ViTaMIn's hardware system overview. The handheld device integrates a GoPro camera for visual information, two tactile sensors for tactile information, a synchronization camera for temporally aligning visual and tactile information, and a rear compartment for storing the battery and a Raspberry Pi. During data collection, the GoPro camera can operate independently, while the two tactile sensors and the synchronization camera connect to the Raspberry Pi in the backbox via USB cables. The Raspberry Pi is powered by the battery inside the backbox, which also houses the data collection control button on its left side. The device is designed for offline operation and is fully functional in the wild. The total weight of the gripper is approximately 1960g. Left: Side view of the ViTaMIn system. Right: Top view of the ViTaMIn system with the backbox cover removed.

tactile representations with reinforcement learning, demonstrating improved data efficiency across diverse manipulation tasks. Similarly, Xu et al. [40] utilized vector-quantized tactile features to improve the performance for robot manipulation and pose estimation.

Aligning heterogeneous sensory modalities is a key challenge in multimodal learning, as different sensors have varying data structures, sampling rates, and noise characteristics [23, 35]. One promising approach to address this challenge is contrastive learning, which maps sensory inputs to a shared latent space, allowing effective cross-modal alignment. Inspired by CLIP [28], researchers have developed contrastive learning techniques that align tactile and visual representations for manipulation tasks [36, 17]. Simeonov et al. [36] leveraged contrastive loss functions to align robot proprioception with vision, improving generalization to unseen tasks. Similarly, Lee et al. [17] demonstrated how self-supervised contrastive learning can effectively align vision and touch for robot grasping. Our work extends these efforts by introducing masked contrastive pre-training, where the tactile encoder learns to reconstruct future occluded visual information, further enhancing multimodal understanding.

III. Visuo-Tactile Manipulation Interface

A. System Overview

We design a handheld gripper to collect visuo-tactile demonstrations without requiring teleoperation on physical robots. Our gripper design is illustrated in Figure 2. The gripper consists of an RGB fisheye wrist camera (GoPro 10) for image observation, two Fin Ray grippers equipped with tactile sensors, a synchronization camera for observation temporal alignment, and a Raspberry Pi 5 with a battery for data recording. The total weight of the gripper is approximately 1960g.

Image Observation To capture comprehensive visual information, we employ a GoPro 10 camera with a 155° field-of-view (FoV) fisheye lens. The camera operates at 60 FPS with a resolution of 2704×2028 pixels and is mounted at the end-effector of our ViTaMIn to ensure consistent visual coverage of the manipulation workspace during demonstration collection and policy deployment.

Tactile Observation In UMI [6], two TPU-printed Fin Ray grippers are used to provide compliance and enhance grasping stability. However, these grippers lack tactile sensing capabilities. In our ViTaMIn, we employ an existing compliant Fin Ray gripper with omnidirectional tactile sensing ability. Figure 3 shows the structure of the gripper. A camera is fixed at the base of the finger with white LEDs for illumination. The

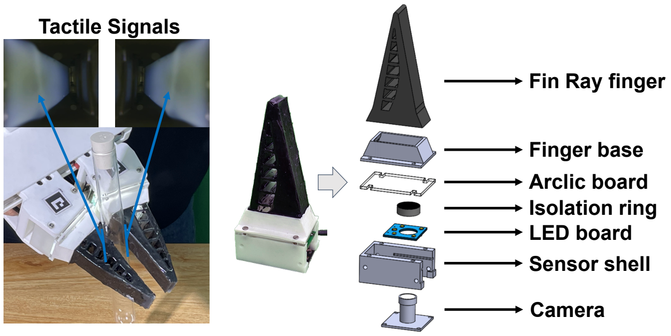

Fig. 3: Exploded view of the tactile sensor structure. The Fin Ray finger is monolithically cast using transparent elastomers, with a semi-transparent layer on the contact surface of the gripper. The finger is then painted black to occlude the environmental illumination.

Fin Ray finger is monolithically cast using transparent elastomers, with a semi-transparent layer on the contact surface of the gripper. The finger is then painted black to occlude the environmental illumination. During manipulation, the camera captures both the global deformation of the entire finger and the local deformation of the contact surface as a single image. The tactile sensing system operates at 30 FPS with a resolution of 640×480 pixels.

Other Observations To enhance the robustness and accuracy of SLAM, we utilize the IMU data provided by the GoPro, which is synchronized with the visual observations. Gripper width is also critical for precise manipulation. Following UMI [6], we attach two ArUco markers to the gripper's fingers and compute the gripper width from the visual observations.

B. Sensor Synchronization

Since the GoPro and the two tactile sensors operate independently, their observations must be synchronized. To achieve this, we use an additional low-cost camera for time alignment. This camera is connected to the Raspberry Pi and is naturally synchronized with the tactile sensors. Before collecting manipulation data, both the synchronization camera and the GoPro simultaneously capture a sequence of ArUco markers displayed on a computer screen. The ArUco IDs are detected in both video streams, and when an identical ID appears in both, the corresponding timestamps are used for synchronization. Since the framerates of the GoPro and the synchronization camera are 60Hz and 30Hz respectively, the temporal alignment error is below $1/60 + 1/30 = 0.05$ seconds, which is sufficient for our tasks. Once the two videos are synchronized, they are cropped by the starting and ending signals triggered by the control button.

C. Data Collection and Filtering

With this design, we adopt a data collection pipeline similar to the one employed in UMI. Similar to UMI, our approach utilizes Simultaneous Localization and Mapping (SLAM) to compute the end-effector poses and uses the delta poses as actions. While SLAM may fail in low-texture environments, it achieves a success rate of approximately 80% in our tasks, allowing the majority of collected data to be used for imitation learning. For details on the number of successful trajectories included, please refer to Appendix E.

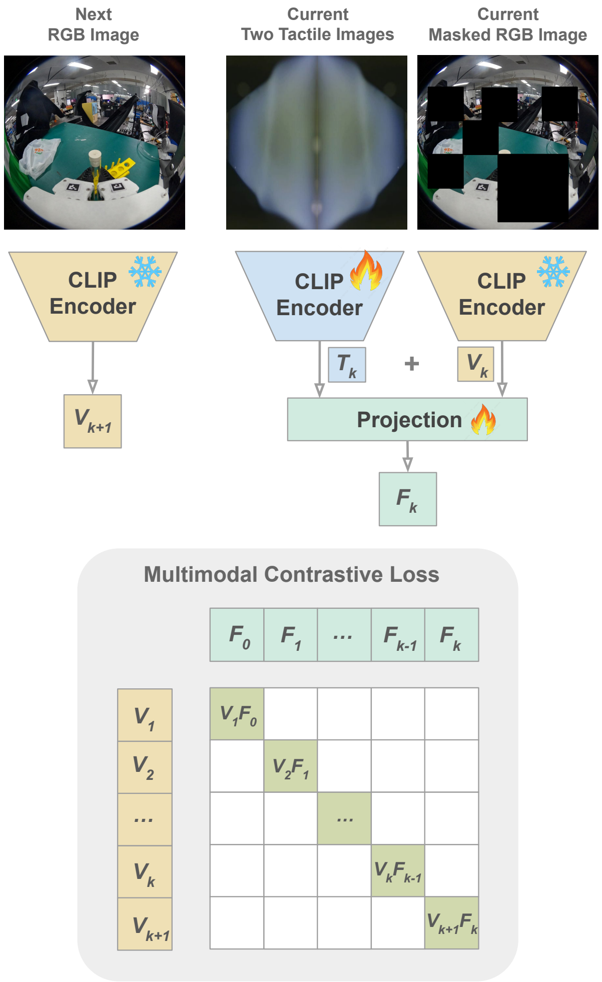

Fig. 4: The illustration of the multimodal representation pretraining approach. The vision encoders are frozen. The tactile encoder is shared for the two tactile images, is initialized from a CLIP ViT-B/16, and is trained to capture complementary information in the tactile image, enabling it to predict the missing content for the future image.

IV. Visuo-Tactile Policy Learning

A. Visuo-Tactile Representation Learning

UMI leverages pre-trained CLIP [28] models to extract image representations. However, ViTaMIn also incorporates two additional tactile images as inputs, which significantly differ from the natural images that CLIP models are originally trained on. As a result, directly applying CLIP models may lead to suboptimal performance due to a mismatch in observation distribution. Meanwhile, ViTaMIn enables the collection of a large dataset, which can be used to pre-train a more effective tactile encoder without relying on the success of SLAM. In this stage, we gather all the collected action-free datasets for the 5 tasks before labeling them with actions, and pre-train an effective tactile encoder that can capture important contact information.

Taking the tactile image in Figure 4 as an example, we want the encoder to capture the essential contact properties, such as the object's in-hand pose and gripper's deformation. These signals are complementary information from pixel observations, and are crucial for making future decisions.

To achieve this, we employ a multimodal contrastive learning approach as illustrated in Figure 4. Given the current masked image $\tilde{I}_V^k$ and current full tactile observation I_T^k of step k, we want the combination of $\tilde{I}_V^k$ and I_T^k align with the future full image observation I_V^{k+1} in the CLIP embedding space. The intuition behind this is to make the tactile encoder focus on the contact information to predict future images based on the current corrupted image.

To ensure stable training, we freeze the image CLIP encoder $\phi_V(\cdot)$ but only fine-tune the tactile encoder $\phi_T(\cdot)$. We first obtain the tactile embedding T_k from $\phi_T(I_T^k)$, and V_k from $\phi_V(\tilde{I}_V^k)$. These embeddings are then concatenated and passed through a fully connected projection layer, mapping them back to the original 512-dimensional CLIP embedding space as a fused feature F_k. Finally, we train the tactile encoder using the standard CLIP loss on F_k and V_{k+1}:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\left(\mathcal{L}_{\text{f-v}} + \mathcal{L}_{\text{v-f}}\right) \tag{1}$$

where

$$\mathcal{L}_{\text{v-f}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(V_{i+1}, F_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(V_{i+1}, F_j)/\tau)} \tag{2}$$

$$\mathcal{L}_{\text{f-v}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(F_i, V_{i+1})/\tau)}{\sum_{j=1}^{N}\exp(\cos(F_i, V_{j+1})/\tau)} \tag{3}$$

here τ is a learnable temperature parameter.

Different from George et al. [10], where they directly apply the CLIP loss on the time-aligned visuo-tactile images I_V^k, we instead fuse the tactile observation with a masked current image to predict the future image. We made this choice for two main reasons. First, in George et al. [10], the tactile representation is conditioned on proprioceptive states, which are unavailable in our dataset before the success of SLAM. Second, since different tasks may have varying images but similar tactile observations, fusing a masked current image helps the network learn a more expressive tactile representation with less confusion. Without sufficient masking, however, the alignment becomes trivial.

B. Behaviour Cloning

After pre-training, we train a Diffusion Policy [5] on the SLAM-filtered data for each individual task. We use the pre-trained tactile encoders and the original CLIP model to extract tactile and visual representations, respectively. Following Chi et al. [5], we use the delta end-effector pose as the action space, and use a Convolutional U-Net [31] as the noise prediction network and apply DDIM [37] to accelerate the inference. For additional training details, please refer to Appendix C.

V. Experiments

A. Experimental Setup

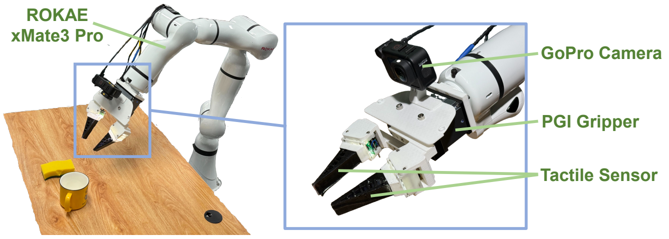

Fig. 5: Experiment setup for policy deployment. The sensing system, including the GoPro camera and the two tactile sensors, is attached to the end effector, in the same configuration as the data collection.

Figure 5 shows the policy deployment setup. We use a Rokae xMate ER3Pro robot arm with a PGI-140-80-W-S parallel gripper fixed at its end. The sensing system, including the GoPro camera and the two tactile sensors, is attached to the end effector, in the same configuration as the data collection.

The system is implemented using ROS Noetic on Ubuntu 20.04. The control loop operates at 10Hz, with separate threads handling visual processing, tactile sensing, and robot control. The system architecture is designed to minimize latency while maintaining reliable real-time performance.

Similar to UMI [6], our system compensates for various sources of latency in the perception-action loop through predictive buffering and timestamp-based synchronization between visual and tactile feedback streams. The policy generates 16 consecutive trajectories at each inference step, with approximately 10 trajectories being executed based on our temporal compensation strategy. For additional deployment details, please refer to Appendix B.

B. Manipulation Tasks

As shown in Figure 6, we design a diverse set of contact-rich manipulation tasks to evaluate the effectiveness of ViTaMIn. These tasks are specifically crafted to demonstrate the following key capabilities: (1) *Robust pick-and-place* of diverse objects, including fragile and small objects; (2) *Dexterous manipulation*, such as in-hand reorientation; (3) *Handling of soft objects* that require adaptive control; (4) *Task success determination*, allowing the robot to repeat attempts until successful completion.

Single-Arm Tasks: We first evaluate our method on 4 single-arm manipulation tasks:

- Orange Placement: The robot is required to pick up a fragile orange from a random initial position and place it onto a randomly placed plate.

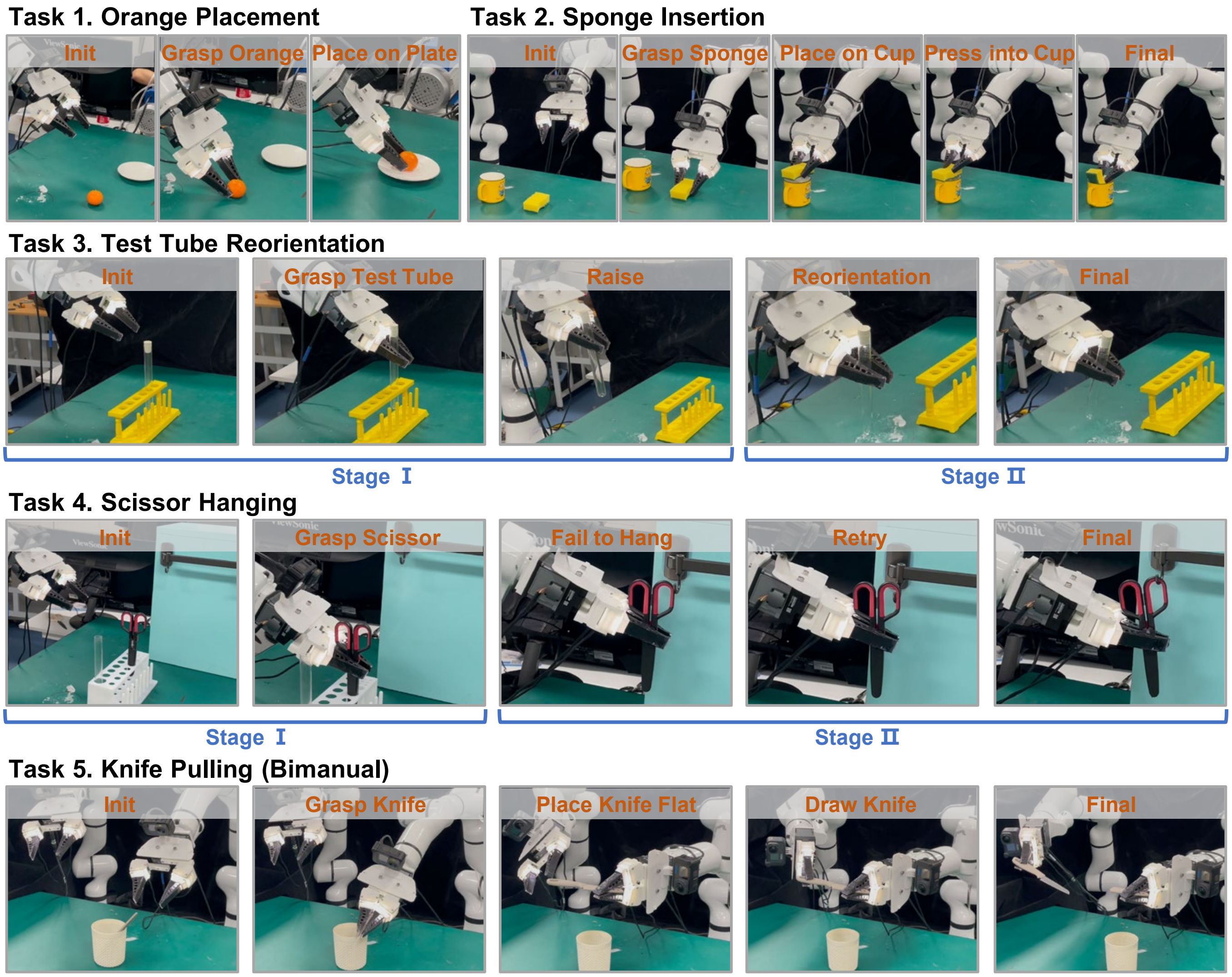

Fig. 6: Task illustration. We test ViTaMIn on 5 contact-rich manipulation tasks. **Orange Placement** tests the ability of pick-and-place of fragile objects. **Test Tube Reorientation** tests the ability of in-hand manipulation guided by tactile sensing. **Scissor Hanging** tests the ability of success determination through multimodal feedback. **Sponge Insertion** tests the ability of precise manipulation for deformable objects. **Knife Pulling** tests the ability of bimanual coordination.

- Test Tube Reorientation: The robot is required to grasp a transparent test tube from a shelf and adjust the test tube's in-hand pose through extrinsic dexterity based on tactile feedback.
- Scissor Hanging: The robot is required to grasp a pair of scissors and hang them on a hook. The robot should be able to adjust the pose and keep attempting until it succeeds.
- Sponge Insertion: The robot is required to first grasp a sponge place it onto a cup with a smaller diameter than the sponge, and then push it to deform so that it can fit into the cup.

Dual-Arm Task: We further evaluate the effectiveness of ViTaMIn on one dual-arm manipulation task, which is Knife Pulling: Two robots are required to coordinate with each other to accomplish this task. The left arm first grasps a knife from a cup, orients it horizontally, and holds it. Then, the right arm reaches for the knife's handle and pulls it out. This task requires both visual information to localize the knife's pose and tactile feedback to grasp the thin object and perform the pulling motion in the correct direction.

We compare our approach against the following methods:

- Vision: For this baseline, the policy only takes visual observation from the GoPro camera. The image is encoded with the pre-trained CLIP model, which is identical to the original UMI [6] paper.
- Ours w/o Pre-training: This baseline integrates visual and tactile observations through simple concatenation after separate ViT-B/16 encoders, which are initialized from the original CLIP model, and fine-tuned during behavior

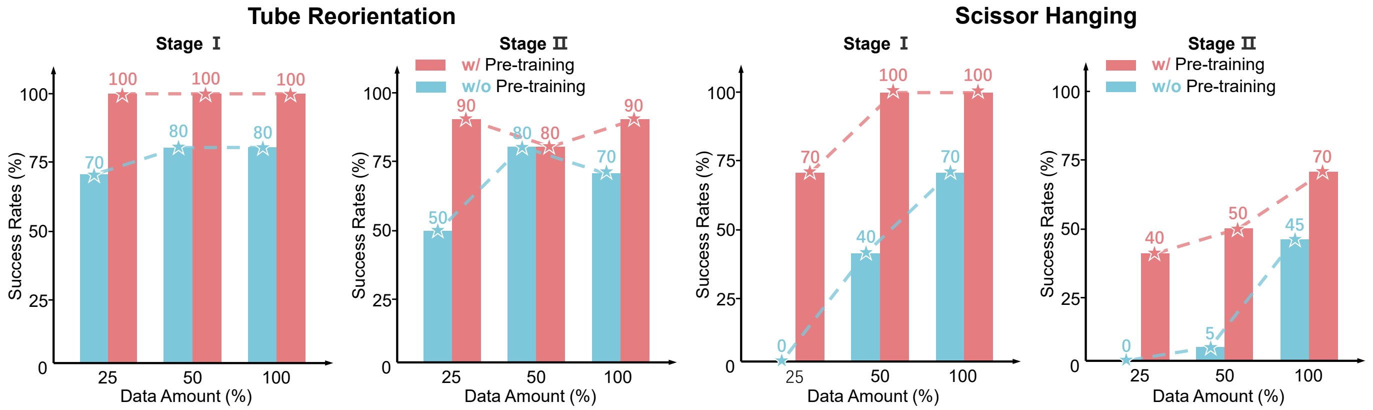

Fig. 7: Ablation study on the effect of pre-training on data efficiency. The performance of the policy improves as the quantity of data increases. When the policy network is pre-trained on the action-free, task-ignorant dataset, our method can achieve a high success rate even with limited data (25%).

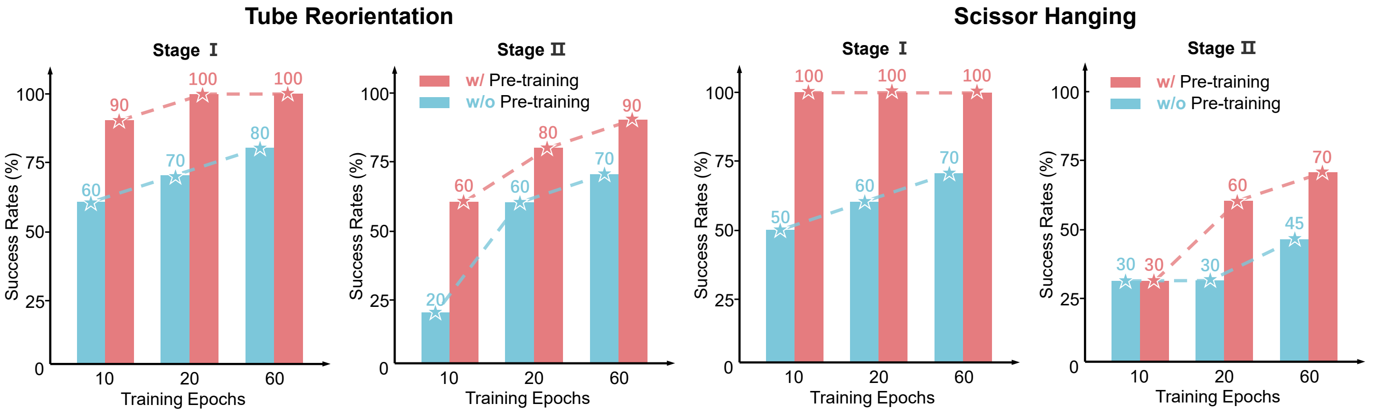

Fig. 8: Ablation study on the effect of pre-training on training efficiency. Policies with pre-training are able to learn to complete the first-stage task at a remarkably early stage of training (within 10 epochs). Additionally, when the policy network is pre-trained, the overall success rates increase more rapidly.

TABLE I: Comparisons on 5 tasks with vision-only policy (original UMI) and ours w/o pre-training to study the effectiveness of tactile signals and multimodal representation learning.The results demonstrate that our approach significantly improves performance across both single-arm and dual-arm tasks.

Task	Vision	Ours w/o Pre-training	Ours
	Single-Arm Tasks		
Orange placement	0.85	0.9	**1**
Test Tube Reorientation	0.4	0.7	**0.9**
Scissor Hanging	0	0.45	**0.7**
Sponge Insertion	0.1	0.4	**0.45**
	Dual-Arm Task		
Knife Pulling	0.6	0.8	**0.9**

cloning with reduced learning rates.

The results are presented in Table I. For each task, we conduct 20 trials with randomized initial conditions and report the average performance. The vision-only policy performs the worst across all five tasks, particularly in contact-rich tasks like test tube reorientation and scissor hanging, where tactile feedback is crucial for success. Across all tasks, pre-training enhances the performance, highlighting the importance of learning effective tactile representations.

C. Ablation Studies

In this section, we evaluate the influence of pre-training on data efficiency and training efficiency.

a) Data Efficiency: We evaluate the performance of policies trained on different amounts (25%, 50%, and 100%) of demonstration data for two tasks. All the models are evaluated in 20 real-world trials with slightly different initial conditions. For a more in-depth analysis, we calculate the success rates of each stage separately, where the stages are illustrated in Figure 6. Figure 7 presents the results. For both methods, the performance of the policy improves with an increase in the quantity of data. With the pre-trained tactile representations, our method can achieve consistently higher success rates on all

TABLE II: Generalization under different test conditions. The results demonstrate that the tactile information and the multimodal pre-training significantly improve the generalization capability to novel objects and different lighting conditions.

Task	Method	Original	Novel Objects	Different Lighting
Orange Placement	Vision	0.85	0.7	0.55
	Ours w/o Pre-training	0.9	0.8	0.6
	Ours	**1.0**	**1.0**	**0.85**
Scissor Hanging	Vision	0.0	0.0	0.0
	Ours w/o Pre-training	0.45	0.4	0.4
	Ours	**0.7**	**0.7**	**0.5**

the tasks across different amounts of data, and can even master the task with limited data (25%) for test tube reorientation.

b) Training Efficiency: We further evaluate the policies trained with different numbers of epochs to understand its training efficiency. All the models are evaluated in 20 real-world trials with slightly different initial conditions. The results are illustrated in Figure 8. Similar to the data efficiency performance, we also observe consistent performance improvements on all the tasks. Moreover, the policies with pre-trained tactile representation can learn to complete the first-stage task at a remarkably early training stage (within 10 epochs). The total success rates also increase faster with pre-training.

D. Generalization Capability

To further investigate the impact of tactile inputs and effective representation learning, we evaluate our policy's generalization to unseen objects and environments, comparing it against its vision-only and pre-training-free counterparts. As shown in Figure 9, beyond the training orange and scissor, we introduce 6 unseen small objects and 3 unseen scissors to assess object generalization. Additionally, we modify lighting conditions by increasing brightness and introducing colored disco ball lighting. Table II presents results on the tasks of orange placement and scissor hanging. We measure the policy's success rate over 20 trials per task and report the average performance.

According to Table II, our method with pre-training achieves consistent better performance across various generalization settings. Moreover, the effect of representation learning also noticeably improves the generalization capability for both tasks.

E. Significance of High-Resolution Tactile Sensing

One key advantage of vision-based tactile sensing over other tactile sensing mechanisms is its high resolution, which is essential for precise manipulation tasks. In this section, we evaluate the significance of high-resolution tactile sensing in two tasks that heavily rely on detailed tactile feedback: Test Tube Reorientation and Scissor Hanging. To demonstrate the importance of high-resolution tactile sensing, we downsample the tactile image to a resolution of 16×16, which is the same resolution as the resistive tactile sensors used in Huang et al. [15]. We measure the policy's success rate over 20 trials per task and report the average performance. Table III presents the results. The high-resolution tactile sensing significantly outperforms the low-resolution counterpart, demonstrating the importance of high-resolution tactile sensing.

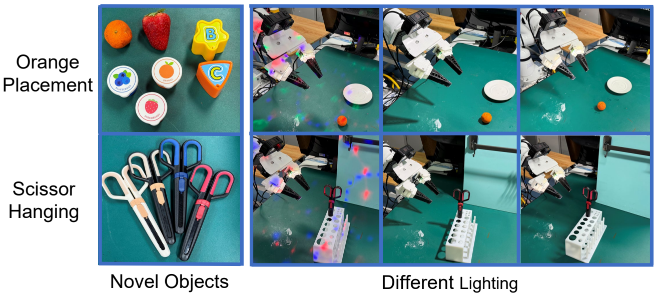

Fig. 9: Showcase of novel objects and different lighting in the generalization task. The left column represents the novel objects. The right three columns illustrate different lighting conditions: colored flashlight illumination, high-power lighting illumination, and normal lighting conditions.

TABLE III: Comparison of different tactile resolution. The results indicate that higher resolution tactile sensing significantly improves the performance for tasks that require fine tactile feedback.

Task	Vision	Vision+ Low-Res Tactile	Vision+ High-Res Tactile
Test Tube Reorientation	0.4	0.6	**0.7**
Scissor Hanging	0	0.2	**0.45**

VI. Limitation and Conclusion

This paper primarily focuses on fixed-base single-arm and dual-arm tasks with parallel-jaw grippers. While this setup is suitable for a wide range of manipulation tasks, it inherently limits the ability to explore more dexterous and contact-rich interactions. Future work could extend our approach to dexterous hands, enabling richer and more versatile manipulation skills that better approximate human-level dexterity. Moreover, long-horizon multi-stage mobile manipulation tasks are likely to gain advantages from multimodal sensing.

In this paper, we present ViTaMIn, a portable visuo-tactile manipulation interface designed for efficiently collecting high-quality demonstrations by capturing both visual and tactile signals. Furthermore, ViTaMIn introduces an effective pre-training strategy that leverages all the collected action-free data to learn a robust and generalizable tactile representation through multimodal contrastive learning. Our approach significantly outperforms vision-only policies across five real-world contact-rich manipulation tasks and demonstrates improved data efficiency, robustness, and generalizability with pre-trained visuo-tactile representations.

References

[1] Jorge Aldaco, Travis Armstrong, Robert Baruch, Jeff Bingham, Sanky Chan, Kenneth Draper, Debidatta Dwibedi, Chelsea Finn, Pete Florence, Spencer Goodrich, et al. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[4] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[5] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

[6] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[7] Kiran Doshi, Yijiang Huang, and Stelian Coros. On hand-held grippers and the morphological gap in human manipulation demonstration. *arXiv preprint arXiv:2311.01832*, 2023.

[8] Hongjie Fang, Hao-Shu Fang, Yiming Wang, Jieji Ren, Jingjing Chen, Ruo Zhang, Weiming Wang, and Cewu Lu. Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15031–15038. IEEE, 2024.

[9] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[10] A George, S Gano, P Katragadda, and AB Farimani. Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. *arXiv preprint arXiv:2403.11898*, 2024.

[11] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

[12] Yunhai Han, Kelin Yu, Rahul Batra, Nathan Boyd, Chaitanya Mehta, Tuo Zhao, Yu She, Seth Hutchinson, and Ye Zhao. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. *IEEE/ASME Transactions on Mechatronics*, 2024.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[14] Koh Hosoda, Katsuji Igarashi, and Minoru Asada. Adaptive hybrid visual servoing/force control in unknown environment. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96*, volume 3, pages 1097–1103. IEEE, 1996.

[15] Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. *arXiv preprint arXiv:2410.24091*, 2024.

[16] Josip Josifovski, Mohammadhossein Malmir, Noah Klarmann, Bare Luka Žagar, Nicolás Navarro-Guerrero, and Alois Knoll. Analysis of randomization effects on sim2real transfer in reinforcement learning for robotic manipulation tasks. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10193–10200. IEEE, 2022.

[17] Michelle A. Lee, Roberto Calandra, Sergey Levine, and Edward H. Adelson. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[18] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[19] Shoujie Li, Haixin Yu, Wenbo Ding, Houde Liu, Linqi Ye, Chongkun Xia, Xueqian Wang, and Xiao-Ping Zhang. Visual–tactile fusion for transparent object grasping in complex backgrounds. *IEEE Transactions on Robotics*, 2023.

[20] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

[21] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023.

[22] PKAAT Miller and PYOBS Leibowitz. Integration of vision, force and tactile sensing for grasping. *Int. J. Intell. Mach*, 4:129–149, 1999.

[23] Anusha Nagabandi, Gregory Kahn, Sergey Levine, and Chelsea Finn. Deep reinforcement learning for vision-based robotic control with multimodal inputs. In *Con-*

ference on Robot Learning (CoRL), 2020.

[24] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 892–909. PMLR, 2022.

[25] Hirofumi Nakagaki, Kosei Kitagaki, Tsukasa Ogasawara, and Hideo Tsukune. Study of deformation and insertion tasks of a flexible wire. In *Proceedings of International Conference on Robotics and Automation*, volume 3, pages 2397–2402. IEEE, 1997.

[26] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, pages 2549–2564. PMLR, 2023.

[27] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dietor Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[29] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023.

[30] Carlo Rizzardo, Sunny Katyara, Miguel Fernandes, and Fei Chen. The importance and the limitations of sim2real for robotic manipulation in precision agriculture. *arXiv preprint arXiv:2008.03983*, 2020.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[32] Felipe Sanches, Geng Gao, Nathan Elangovan, Ricardo V. Godoy, Jayden Chapman, Ke Wang, Patrick Jarvis, and Minas Liarokapis. Scalable. intuitive human to robot skill transfer with wearable human machine interfaces: On complex, dexterous tasks. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6318–6325, 2023. doi: 10.1109/IROS55552.2023.10341661.

[33] Carmelo Sferrazza, Younggyo Seo, Hao Liu, Youngwoon Lee, and Pieter Abbeel. The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9698–9705. IEEE, 2024.

[34] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[35] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, and Dieter Fox. Perceiver for robotics: Generalized multimodal perception with transformer architectures. *Robotics: Science and Systems (RSS)*, 2023.

[36] Anthony Simeonov, Suraj Nair, and Chelsea Finn. Neuralskill: Neural representations for efficient skill learning and adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[38] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv:2203.06173*, 2022.

[39] Pengwei Xie, Rui Chen, Siang Chen, Yuzhe Qin, Fanbo Xiang, Tianyu Sun, Jing Xu, Guijin Wang, and Hao Su. Part-guided 3d rl for sim2real articulated object manipulation. *IEEE Robotics and Automation Letters*, 2023.

[40] Zhengtong Xu, Raghava Uppuluri, Xinwei Zhang, Cael Fitch, Philip Glen Crandall, Wan Shou, Dongyi Wang, and Yu She. UniT: Unified tactile representation for robot learning, 2024. URL https://arxiv.org/abs/2408.06481.

[41] Xinwei Zhang and et al. Fusing multimodal sensory data for robotic perception. *IEEE Transactions on Robotics*, 2022.

[42] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Appendix

A. Comparison of Grippers

Figure 10 compares the TPU-printed Fin Ray gripper employed in UMI [6] with the elastomer-casted Fin Ray gripper used in our work. Our gripper is softer and can offer greater compliance and caging capability. Since the TPU-printed finger is more rigid and lacks tactile feedback, it may damage fragile objects such as strawberries during grasping. Another advantage of our tactile sensors is their ability to deform and sense contact in all directions. Thus, when in contact with the table, the sensor can deform, detect the contact, and respond appropriately. This is particularly crucial for grasping small objects on the table when the SLAM accuracy and the trained policy are not precise enough.

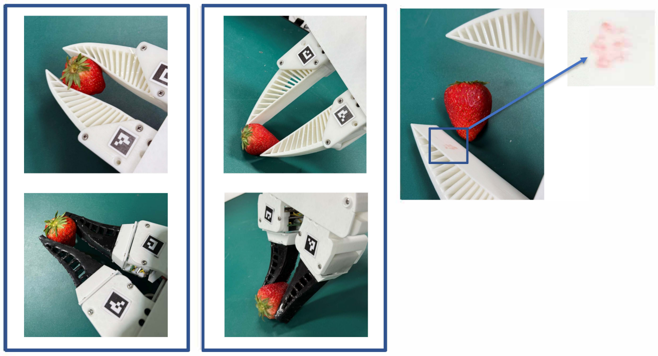

Fig. 10: Comparison of the TPU-printed Fin Ray gripper employed in UMI [6] with the elastomer-casted Fin Ray gripper used in our work. The strawberry is damaged by the TPU gripper during grasping.

B. Implementation Details

A. Hardware Setup

Our system consists of a Rokae xMate ER3PRO robotic arm equipped with a PGI-140-80-W-S parallel gripper. The 7-DOF robotic arm provides flexible manipulation capabilities, while the gripper features an 8cm stroke range from fully open to closed position. For visual perception, we employ a GoPro 10 camera with a 155° field-of-view fisheye lens, operating at 60 FPS with a resolution of 2704×2028 pixels. The tactile sensing system consists of two Fin Ray grippers, each integrated with a camera operating at 30 FPS with a resolution of 640×480.

As illustrated in Figure. 3 of the main text, our gripper design incorporates several key components. The Fin Ray finger is fabricated through a single-piece casting process using transparent silicone. The finger base and sensor shell are 3D printed using photocurable resin, serving as mounting points for the gripper and camera respectively. An acrylic board provides both light transmission and structural support. The LED board illuminates the blackened gripper surface, enabling the camera to capture internal deformations. An isolation ring prevents direct light exposure from the LED to the camera.

B. Software Setup

Our system is implemented using ROS Noetic on Ubuntu 20.04. The control loop operates at 10Hz, with separate threads handling visual processing, tactile sensing, and robot control. The system architecture is designed to minimize latency while maintaining reliable real-time performance.

Our system is deployed on a PC with an NVIDIA RTX 2080Ti GPU. The tactile sensors are directly connected to the workstation via USB cables, while the GoPro camera feed is captured through an Elgato HD60 X capture card with an external media module.

C. Latency Compensation

Similar to UMI [6], our system compensates for various sources of latency in the perception-action loop through predictive buffering and timestamp-based synchronization between visual and tactile feedback streams. The policy generates 16 consecutive trajectories at each inference step, with approximately 10 trajectories being executed based on our temporal compensation strategy.

C. Training Details

Our policy learning implementation and training are largely based on Diffusion Policy [5] and UMI [6]. For all the experiments of ours and baselines, we use consistent policy training hyper-parameters, as shown in Table V. Table IV shows the multimodal representation learning hyperparameters.

For all the experiments, we evaluate the checkpoints at 60 training epochs after convergence at default.

TABLE IV: Hyper-parameters for representation learning.

Parameter	Value
Observation resolution	224×224
Mask ratio	$[0.5, 0.75]$
Optimizer	AdamW
Optimizer momentum	$\beta_1 = 0.95$, $\beta_2 = 0.999$
Learning rate	$1e{-}4$
Learning rate schedule	Cosine decay
Batch size	256

TABLE V: Training hyper-parameters for policy learning.

Parameter	Value
Observation Settings	
Image observation horizon	2
Proprioception observation horizon	2
Action horizon	16
Observation resolution	224×224
Optimization Parameters	
Optimizer	AdamW
Optimizer momentum	$\beta_1 = 0.95$, $\beta_2 = 0.999$
Learning rate (action diffusion)	$3e{-}4$
Learning rate (visual encoder)	$3e{-}5$
Learning rate schedule	Cosine decay
Batch size	64
Train diffusion steps	50
Inference denoising steps	16

D. Task Description and Failure Analysis

We evaluate our system on four single-arm tasks and one dual-arm task.For each task, we evaluate the system's performance across 20 different initial configurations, varying both the robot's initial end-effector position and the target object's placement.

1) Orange placement Task: The orange and plate are placed at random initial positions within a 50cm×50cm workspace area on the table. The task requires the gripper to securely grasp the orange and place it on the plate. Success is determined by stable placement of the orange on the target plate.

Failure Modes:

- Policy trajectory generation failures despite clear visual detection
- Gripper collision with the table surface during grasping
- Unsuccessful placement after successful grasping

Vision baselines particularly struggle with table collision avoidance.

2) Test tube reorientation: The test tube is randomly placed in one of the holes of a tube rack, with the rack position varying within a 20cm×20cm area. The robot needs to grasp the tube and reorient it to a vertical position. The task is considered successful when the tube's orientation error is less than 10° from vertical.

Failure Modes:

- Collisions with the tube rack during grasping
- Excessive lifting without proper table-relative positioning
- Incorrect reorientation of the tube

Vision approaches primarily fail in orientation assessment due to lack of tactile feedback.

3) Scissor hanging: A pair of scissors is placed on a rack within a 20cm×20cm workspace area. The robot needs to pick it up and hang it on a hook mounted on a vertical surface, with approximately 1.5cm clearance between components. Success is defined by stable hanging and successful gripper release within five retry attempts.

Failure Modes:

- Unsuccessful scissor detection and localization
- Failed hanging attempts after successful grasping
- Failure to release after successful hanging

Vision methods struggle with release timing due to lack of tactile feedback for task completion detection.

4) Sponge insertion: The sponge (15cm×8cm×2cm) and cup (4.5cm inner radius) are randomly placed within a 30cm×30cm area on the table. Success is determined by the sponge making contact with the bottom of the cup.

Failure Modes:

- Misaligned grasping of the sponge
- Unsuccessful placement of the sponge on the target surface
- Deformation control during insertion

Vision approaches have difficulty managing the deformable nature of the sponge.

5) Knife pulling: This dual-arm task requires coordinated motion between the two arms. The knife is placed in a knife holder, with the holder's random range being 15cm×15cm. The left arm grasps and orients the knife to a horizontal position, while the right arm grasps the handle and performs a pulling motion. The task is considered successful once the knife has been fully pulled out.

Failure Modes:

- Inter-arm collisions due to the knife's small form factor
- Imprecise positioning of the right arm
- Loss of grasp during coordinated motion

Vision methods particularly struggles with maintaining stable grasps during the coordinated pulling motion.

Figure 11 shows some representative failure cases.

E. Demonstration Data Statistics

TABLE VI: Data Collection Statistics for Different Tasks

Task	Raw Data	Valid Data*	Avg. Length
Orange Placement	87	73	435
Test Tube Reorientation	150	125	619
Sponge Insertion	160	138	605
Scissor Hanging	172	137	642
Knife Pulling (Left)	188	131	403
Knife Pulling (Right)	180	134	254

*Valid data refers to demonstrations with successful SLAM tracking

Table VI shows the statistics of the demonstration data. We collecte demonstrations for both single-arm and dual-arm manipulation tasks. For single-arm tasks, we gather between 87 and 172 raw demonstrations per task according to the task difficulty, with successful SLAM tracking achieved in approximately 80% of the trajectories. The dual-arm knife pulling task requires coordinated motion between both arms, with similar data collection volumes but slightly different average demonstration lengths for left and right arm movements.

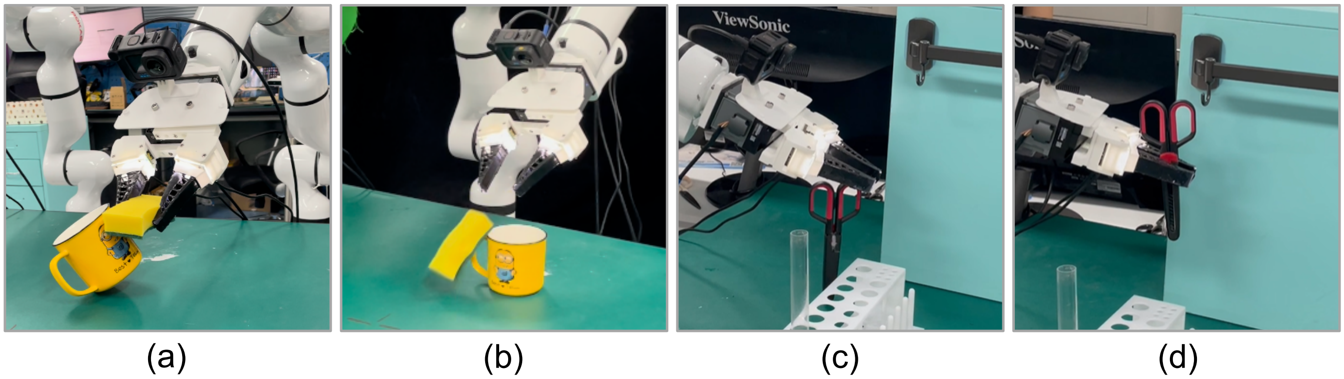

Fig. 11: Representative failure modes in Sponge Insertion and Scissor Hanging. (a) The gripper knocks over the cup. (b) The sponge is not successfully placed on the cup. (c) The gripper fails to tightly grasp the scissor, causing it to fall. (d) The scissor is still not successfully hung after multiple retries.