

Gentle Manipulation Policy Learning via Demonstrations from VLM Planned Atomic Skills

Jiayu Zhou^{1*}, Qiwei Wu^{2*}, Jian Li², Zhe Chen¹, Xiaogang Xiong^{1†}, Renjing Xu^{2†}

¹Harbin Institute of Technology (Shenzhen)

²Hong Kong University of Science and Technology (Guangzhou)

Abstract

Autonomous execution of long-horizon, contact-rich manipulation tasks traditionally requires extensive real-world data and expert engineering, posing significant cost and scalability challenges. This paper proposes a novel framework integrating hierarchical semantic decomposition, reinforcement learning (RL), visual language models (VLMs), and knowledge distillation to overcome these limitations. Complex tasks are decomposed into atomic skills, with RL-trained policies for each primitive exclusively in simulation. Crucially, our RL formulation incorporates explicit force constraints to prevent object damage during delicate interactions. VLMs perform high-level task decomposition and skill planning, generating diverse expert demonstrations. These are distilled into a unified policy via Visual-Tactile Diffusion Policy for end-to-end execution. We conduct comprehensive ablation studies exploring different VLM-based task planners to identify optimal demonstration generation pipelines, and systematically compare imitation learning algorithms for skill distillation. Extensive simulation experiments and physical deployment validate that our approach achieves policy learning for long-horizon manipulation without costly human demonstrations, while the VLM-guided atomic skill framework enables scalable generalization to diverse tasks.

Introduction

Robotic manipulation has made significant progress, yet increasingly complex tasks demand both long-horizon continuous operation and fine-grained, contact-rich interactions. Traditional methods primarily rely on coarse visual perception, which falls short in capturing the delicate contact dynamics essential for precise manipulation (Chi et al. 2023a; Ze et al. 2024). Tactile sensing offers high-resolution surface and contact state information critical for enabling gentle and accurate physical interactions (Xue et al. 2025a; Huang et al. 2024). As manipulation tasks evolve from short, discrete stages to extended sequences requiring sustained dexterity and adaptability (Shi et al. 2023), addressing temporal dependencies and ensuring safe, gentle contact over prolonged periods becomes increasingly challenging.

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Currently, long-horizon gentle manipulation remains a challenge. In imitation learning-based methods, the acquisition of expert demonstration data remains a major bottleneck due to its high cost and labor intensity. While visual-language model frameworks have shown promise in automating task planning and policy learning, current approaches largely overlook the integration of tactile and contact force information. The difficulties in achieving manipulation tasks through these methods can be summarized as follows. First, most existing public data benchmarks lack consideration and focus on force, yet in practical tasks, it is necessary to avoid damaging objects during manipulation. Second, perceptual uncertainty arising from sensor noise and partial observability complicates reliable decision-making. Finally, current data acquisition methods are expensive, and there are no effective ways to synthesize and expand data.

In this work, we propose a force-aware synthetic data generation approach through simulation-trained tactile atomic skills, VLM-guided hierarchical task decomposition, and the generation of multimodal long-horizon manipulation data, establishing a benchmark methodology for imitation learning to address these challenges, as shown in Fig. 1. Our main contributions include:

- **Force-Aware and Scalable Data Generation:** We present a long-horizon sequential manipulation data generation framework that explicitly incorporates contact force information, improving reliability and safety while substantially reducing manual effort in constructing expert demonstrations. Leveraging a library of robust atomic skills, the framework supports scalable extension to diverse complex tasks through hierarchical composition.
- **Dataset and Training Benchmark VT-DP:** Based on our data generation method, we have prepared a small, scalable dataset with language labels that is based on visual-tactile 3D point clouds, and provided a benchmark VT-DP (Visual-Tactile Diffusion Policy) capable of enabling multi-modal imitation learning training.

We conducted ablation studies to demonstrate the functionality of each component in our data synthesis framework. Additionally, we presented comparisons with different data collection methods, comparisons with different training

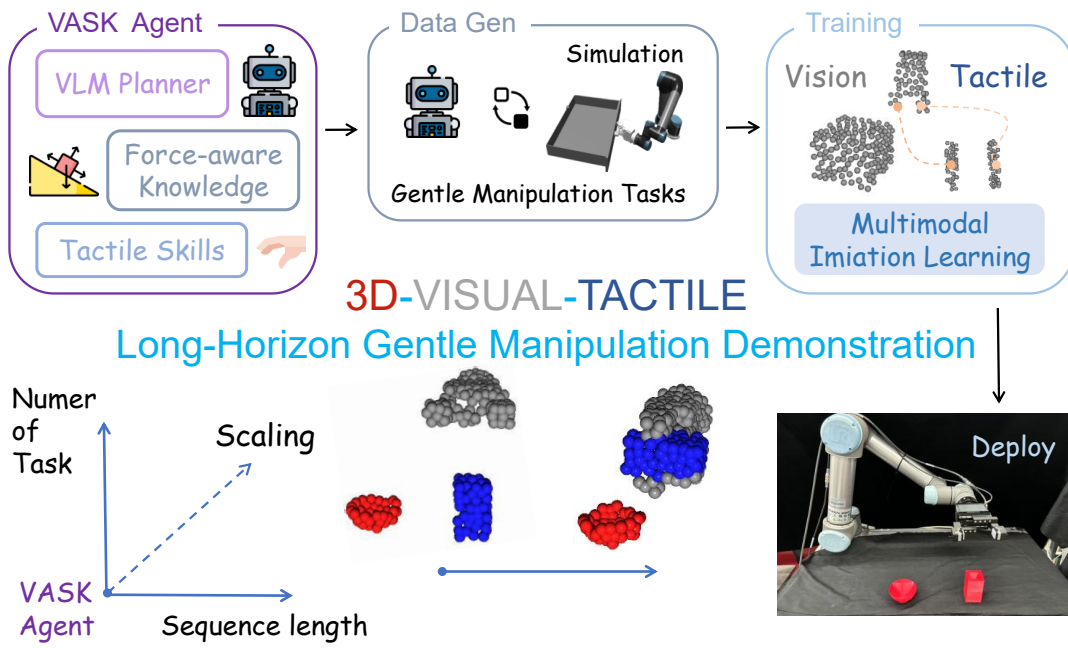


Figure 1: Our pipeline trains robotic arm tactile skills through force-constrained reinforcement learning in simulation. Visual Language Models then plan task sequences by interpreting visual scenes and language instructions to generate expert demonstrations. These demonstrations are distilled into contact-aware manipulation policies via visual-tactile Diffusion Policy, enabling end-to-end execution of long-horizon tasks from multi-modal point cloud inputs.

frameworks, and successful policy deployment in the real world. Experimental results demonstrate that the final policy of our framework achieves superior performance, robust generalization capability, and significant extensibility.

Related Work

Contact-Rich Long-Horizon Manipulation

In robotic manipulation, traditional control methods typically require accurate dynamic models to perform complex tasks (Li et al. 2024; Jin and Posa 2022; Jiang et al. 2024). However, obtaining precise models is often infeasible in real-world scenarios due to model uncertainty and environmental variability. Reinforcement learning algorithms (Liu et al. 2023; Chebotar et al. 2023) have demonstrated the ability to learn manipulation policies through repeated interaction with the environment without relying on explicit models. Despite these advances, most RL research has focused on single-stage tasks such as grasping and pushing (Zhong et al. 2025; Yang et al. 2023; Wu et al. 2024; Welle et al. 2023). For contact-rich tasks, studies often emphasize peg-in-hole insertion and in-hand manipulation (Ankile et al. 2024; Geng and Liu 2023), which remain relatively constrained problems. As manipulation tasks become more complex and long-horizon in nature, designing appropriate reward functions becomes increasingly challenging, making it difficult for RL agents to converge to stable and effective policies. To address this, imitation learning (Xue et al. 2025b; Chi et al. 2023a; Zhao et al. 2023; Ze et al. 2024) has been widely adopted for long-horizon tasks by leveraging

expert demonstrations to guide the agent. Nevertheless, imitation learning approaches typically require extensive and costly data collection from manual demonstrations, limiting scalability. Unlike prior works that rely exclusively on reinforcement or imitation learning, our framework leverages the strengths of both. We first train robust single-stage tactile atomic skills via reinforcement learning in simulation. Complex long-horizon tasks are then decomposed into sequences of these primitives, enabling the generation of high-quality expert demonstration data. Finally, we distill these demonstrations through imitation learning into a stable, gentle manipulation policy capable of long-horizon execution. This hybrid approach reduces data collection and training burdens while improving adaptability and performance in challenging, contact-intensive long-horizon manipulation tasks.

Visual Language Models for Data Generation

Integrating VLMs into robotic systems (Chen et al. 2025; Yang et al. 2025; Kim et al. 2024) significantly enhances intelligence and generalization by translating natural language into executable action sequences. Recent approaches leverage foundation models to automatically generate diverse tasks and scenarios, scaling robot datasets with minimal human input. RoboGen (Wang et al. 2023b) utilizes language descriptions to generate scene configurations and learns skills through LLM-synthesized reward functions. GenSim (Wang et al. 2023a) employs LLMs for procedural generation of scenes, simulations, and demonstrations, while GenSim2 (Hua et al. 2024) extends this capability to long-horizon articulated tasks through enhanced reason-

ing models. HumanoidGen (Jing et al. 2025) further applies this paradigm to bimanual dexterous manipulation using pre-trained sub-policies and annotated 3D assets. However, these existing methods generally do not incorporate contact force information. In contrast, our framework explicitly integrates contact force data by leveraging pretrained atomic policies trained with force-aware reward functions combined with VLM-based compositional planning. This novel integration not only facilitates stable and physically consistent low-level execution but also enables the automatic generation of high-quality demonstrations for complex long-horizon manipulation tasks involving rich physical interactions.

Method

Skills for Gentle Manipulation

Contact Force Characterization Contact force is fundamental to robotic manipulation, especially in tasks requiring fine-grained interaction. Typically, contact force can be decomposed into three primary components: normal force, shear force, and torque, each exhibiting distinct characteristics during different manipulation phases. During grasping, the focus lies on the normal force generated by tactile sensors compressing the object’s surface. When the gripper moves the object, the resulting contact forces comprise a combination of normal force, shear force, and torque due to the multidirectional nature of the movement. This classification facilitates targeted analysis and optimization of the robot-object interaction for specific tasks, thereby enhancing both safety and operational efficiency.

Force-Aware Atomic Skills As shown in Fig. 2, leveraging the categorization of contact forces, we designed atomic skills that align with interaction patterns such as “grasp”, “rotate”, and “move”, with each skill focusing on specific force components. To accomplish more complex manipulation tasks, such as opening drawers and boxes, we developed the skill “Horizontal pull”, which primarily involves clamping force and horizontal friction, and the skill “Lateral move”, which combines vertical grasping force with unidirectional pushing force. Together, these atomic skills encompass the diverse contact force distributions encountered in real-world operations, thereby enabling robust task execution.

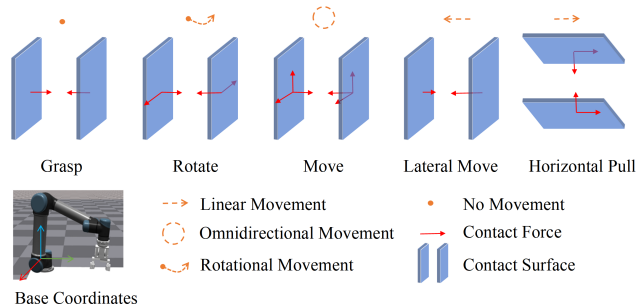


Figure 2: Schematic diagram of contact force of tactile atomic skills.

Implementation of Atomic Skills Reinforcement learning (RL) provides significant advantages for acquiring atomic skills in contact-rich manipulation: its trial-and-error paradigm enables autonomous policy optimization through environmental interaction while accommodating stochastic task dynamics. To leverage these benefits, we implement atomic skill training using the Soft Actor-Critic (SAC) algorithm (Haarnoja et al. 2018). SAC incorporates the maximum entropy principle to encourage policy stochasticity—critical for robust contact interactions—while its twin Q-network architecture reduces value estimation bias to enhance training stability. To further improve policy robustness, we apply domain randomization by dynamically perturbing key physical parameters during training, including object pose initialization, mass, and friction coefficients. All training is performed within the Isaac Gym simulation environment under full state observability. The observation space includes the robot state, object state data, and tactile sensor contact forces. The action space consists of relative Cartesian commands to the end effector, denoted as $\{a_x, a_y, a_z, a_{rx}, a_{ry}, a_{rz}, a_c\}$, where translational (a_x, a_y, a_z) and rotational (a_{rx}, a_{ry}, a_{rz}) components operate along the Tool Center Point (TCP) axes, and a_c controls the relative position of the width of the parallel gripper. Each atomic skill is associated with a specifically designed reward function tailored to its single-stage objective, which incorporates a force penalty term to encourage gentle manipulation policies. For example, the reward function for the grasping skill is formulated as follows:

$$R_{Grasp} = T - \tanh(D_e) + PC_p + H_r + C_r + F_p + S_r \quad (1)$$

where T denotes a time penalty to encourage task efficiency, and the hyperbolic tangent function $\tanh(\cdot)$ normalizes the distance D_e between the gripper fingers and the target object to ensure smooth reward variation. PC_p penalizes premature gripper closure, while C_r incentivizes maintaining contact with the object. H_r rewards the height of the lifted object, F_p penalizes excessive contact force, and S_r provides a large positive reward upon successful task completion.

Generation of Gentle Manipulation Data

Planning with Atomic Skills Constructing demonstration datasets for long-horizon manipulation tasks typically requires substantial manual effort and incurs high costs. To address this challenge, we propose a novel framework, Visual Language with Atomic Skills (VASK), which employs a pre-trained VLM as the top-level planner. We design a comprehensive knowledge base for the VLM that includes an operational context introduction, an atomic skill library, and system prompts. The atomic skill library encompasses various tactile skills, along with detailed descriptions of each skill and their corresponding contact force information during manipulation. For example, the skill “Grasp” involves “Move the gripper to the object and perform a grasp, considering normal contact forces to avoid damaging the object.” During the planning phase, based on a human-provided task description and operational background image, VASK decomposes the task into corresponding sequences of atomic

skills. For example, given the task description: “Open the drawer and place the object inside”, the VLM first analyzes the drawer-opening operation: “This operation involves grasping the drawer handle and pulling it horizontally”. Then it reasons: “The skill Horizontal pull, which considers normal contact force and lateral friction, precisely matches this scenario”. Therefore, the first selected action is “Horizontal Pull”. Subsequently, according to the following task requirements, the model sequentially selects appropriate tactile skills, resulting in an action sequence such as [“Horizontal Pull”, “Grasp”, “Move”, “Open”]. VLM explicitly accounts for the types of contact forces involved in each subtask and searches the atomic skill library to find corresponding tactile skills, thereby planning a reasonable and effective manipulation strategy.

VASK for Data Generation As illustrated in Fig. 3, the complete data generation process comprises four stages: **Initialization**: The manipulation environment is initialized, and the VLM reads information from the knowledge base to guide hierarchical decomposition. **Planning**: Generating hierarchical skill sequences via VLM by integrating natural language instructions and background image. **Execution and Collecting**: The robot sequentially executes atomic skills while continuously acquiring environmental information and contact force feedback, simultaneously recording raw point cloud data during the manipulation process. **Data Processing**: Select trajectories that meet the task completion criteria and exhibit relatively low contact forces to ensure high-quality data and promote gentle manipulation. It is worth noting that our approach uniquely incorporates contact force information into the long-horizon task data generation framework while significantly reducing the manual cost required to construct expert demonstration datasets.

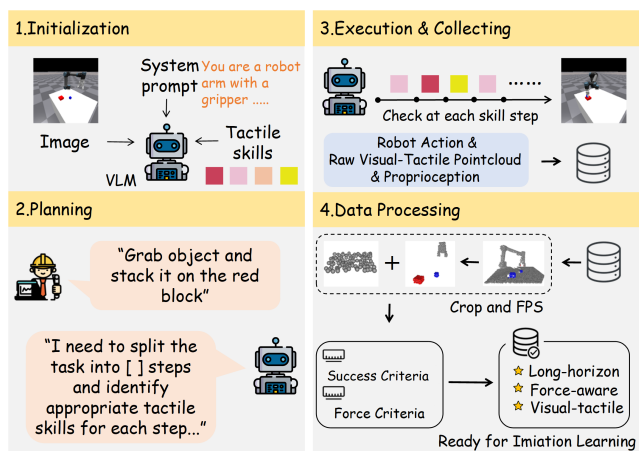


Figure 3: Framework diagram of VASK. The VLM receives system prompts, natural language task descriptions, and RGB images, and integrates this information with atomic skills to guide the agent in completing manipulation tasks and collecting raw trajectory data.

Learning Long-Horizon Gentle Manipulation

The VASK framework is capable of completing long-horizon gentle manipulation tasks, largely benefiting from full state observability available in the simulation environment. However, as task complexity increases, the inherent hallucination tendencies of the VLM lead to instability in the policy outputs. To address this, we employ knowledge distillation to extract operational policies grounded on visuo-tactile point clouds and proprioceptive sensing.

Fusion of Visual and Tactile Modalities The fusion of visual and tactile modalities enables robotic systems to integrate global operational awareness with fine-grained contact details, facilitating complex manipulation tasks. Compared to RGB images providing limited 2D information, point clouds offer robust spatial geometric representations exhibiting strong invariance to illumination variations, background clutter, and color changes. In simulation, visual point clouds are acquired from depth images captured by fixed-mount depth cameras. For tactile point clouds, we adopt the soft contact model from TacSL (Akinola et al. 2024) to simulate contact interactions. In this model, objects are still modeled as rigid bodies but are allowed to mutually penetrate in proportion to the interaction forces. Depth cameras capture these penetration states between rigid bodies to construct tactile point clouds that represent the contact geometry. Tactile point clouds deform in response to contact.

VT-DP for Knowledge Distillation Diffusion models offer substantial advantages for imitation learning in robotic control, as they effectively capture complex, multimodal action distributions and ensure stable training through a gradual denoising process. Prior work DP3 (Ze et al. 2024) demonstrates that leveraging point clouds significantly enhances policy generalization in unstructured environments, making them particularly suitable for diffusion-based policy learning. Building upon these insights, we propose Visual-Tactile Diffusion Policy (VT-DP), a long-horizon gentle manipulation distillation framework that integrates visual and tactile point clouds. Building upon the DP3 method, We crop the global point clouds to retain only the end-effector gripper and the manipulated object, and downsample the visual point cloud to 256 points as well as the tactile point cloud to 128 points. Furthermore, visual and tactile point clouds are encoded separately to better capture modality-specific features. Based on the VT-DP framework, only 50 high-quality demonstration trajectories are sufficient to distill gentle manipulation policies for corresponding long-horizon sequential tasks.

Experiment

Experiment Setup

We design a comprehensive set of experiments to evaluate the performance of our framework, focusing on two key research questions: (1) Whether the distilled policy derived from our framework enables gentle manipulation; (2) The stability and robustness of the manipulation strategies produced by our framework;

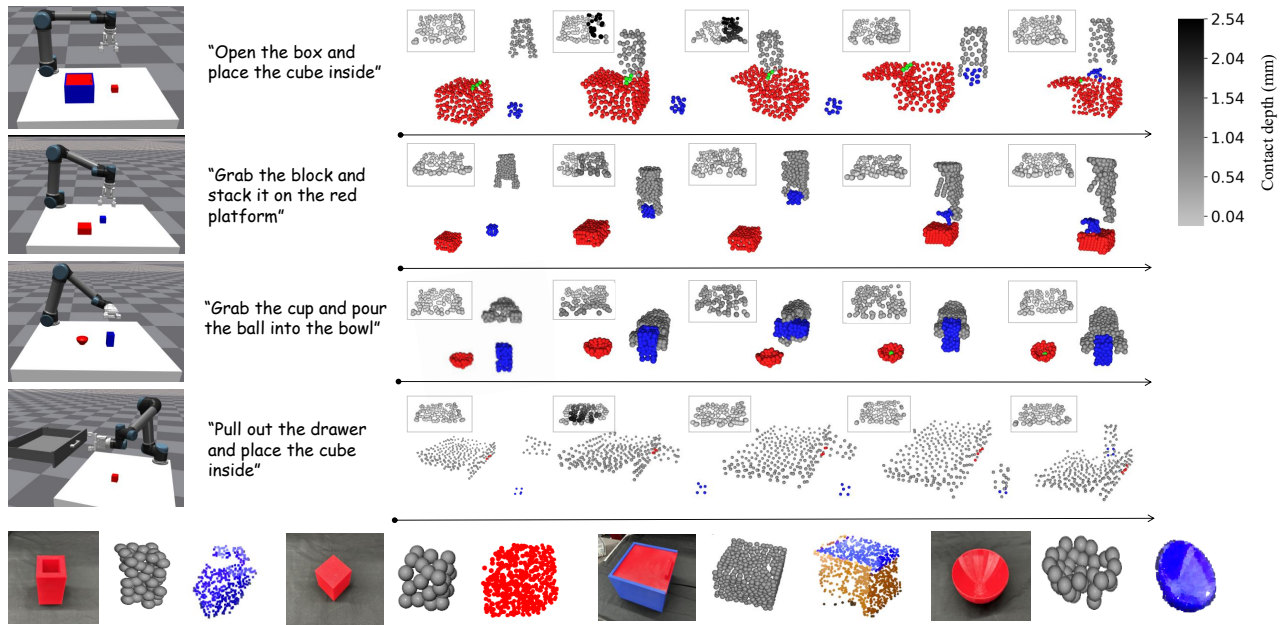


Figure 4: Point cloud sequence during a long-horizon manipulation task. The boxed region in the upper-left corner of the visual point cloud shows the tactile point cloud. The legend indicates the contact depth between the sensor and the object. Note that the point cloud data itself is uncolored; colors are applied solely for visualization purposes.

Robot System Setup Our experimental platform consists of a UR5 robotic arm equipped with two Gelsight Mini tactile sensors and a parallel gripper. We custom-designed adapters to integrate the tactile sensors onto the parallel gripper, enabling comprehensive acquisition of contact information between the gripper and manipulated objects. Additionally, an Intel RealSense D415 RGB-D camera provides visual perception. We constructed the corresponding robotic manipulation system within the Isaac Gym simulation environment and utilized the TacSL tactile sensor model to obtain tactile point clouds. This setup minimizes the discrepancy between simulated and real tactile feedback, thereby facilitating the deployment of learned policies in real-world scenarios.

Benchmark Setup In the simulation environment, we design four manipulation tasks representing diverse challenges. Below is a brief introduction to these tasks:

- **Object Stack (OS):** Generate a random block and a platform. Grasp the block and stack it onto the platform.
- **Open and Place (OP):** Grasp the handle of a box and open it. Move to an object and place the object inside the box.
- **Cabinet and Place (CP):** Move the robot arm to the drawer and open it. Then, move to the object, grasp it, and place it inside the drawer.
- **Pour Ball (PB):** Grasp a cup containing a small ball, move it above a bowl, and rotate the cup to pour the ball into the bowl.

As illustrated in Fig. 4, these long-horizon tasks encompass a variety of fundamental skills required in real-world scenarios. Notably, all tasks share the same atomic skill library.

By flexibly constructing and combining these atomic skills, manipulation data of arbitrary sequence length can be generated, which highlights the scalability and adaptability of our framework across diverse manipulation scenarios.

Training Setup and Model Preparation We record the success rates of each atomic skill as shown in Tab. 1. All skills demonstrate high performance due to their single-stage contact manipulation characteristics. This provides a strong guarantee for generating long-horizon gentle manipulation data. Our VT-DP utilizes a Transformer-based noise prediction network with approximately 13.27 million parameters. Training involves 100 denoising steps to progressively reconstruct action sequences from noisy expert trajectories, while inference employs a reduced 10-step denoising schedule to balance efficiency and performance. The policy operates on an 8-step horizon, consuming 2-step observations to predict 5-step future actions. Inference runs at 28 Hz on an RTX 3060 GPU, executing two actions at each step to ensure real-time responsiveness and task success. Training converges after approximately 800 epochs.

Grasp	Rotate	Move	Lateral move	Horizontal pull
0.98	0.96	0.89	0.92	0.95

Table 1: Success Rates of Gentle Atomic Skills Trained via Reinforcement Learning in Simulation.

Ablations

VLM-Planner Ablation We investigated the combination of different VLMs with atomic skills to demonstrate the

generalizability of our framework. For each VLM, we utilized pretrained weights directly without any task-specific fine-tuning. The system prompt, environment RGB images, and relevant task descriptions were kept consistent across experiments. Our evaluation focuses on VLM accuracy in atomic skill planning, and transition point judgment. We recorded the success rates of trajectory generation across multiple tasks, with the experimental results summarized in Tab. 2. Among the tested VLMs, the Qwen2.5-vl-7b-instruct model exhibited significantly lower success rates, likely due to its smaller model size and limited instruction-following capability, which hindered accurate planning and skill transitions. In contrast, Doubao-vision-pro demonstrated superior performance in both atomic skill planning and transition point detection. Considering its strong performance and ease of use, we adopt the Doubao-vision-pro model combined with atomic policies to generate expert demonstrations within the VASK framework.

VLM Model	OS	OP	CP	PB	Avg
Doubao-pro	0.96	0.81	0.5	0.58	0.713
Doubao-lite	0.19	0.20	0.22	0.13	0.185
Qwen-max	0.96	0.55	0.67	0.50	0.670
Qwen2.5-7b	0.13	0.00	0.00	0.00	0.033
Qwen2.5-32b	0.86	0.65	0.56	0.28	0.588
Qwen2.5-72b	0.91	0.54	0.51	0.47	0.608

Table 2: Success Rates of Trajectory Generation for Different VLM Models and atomic Skill Compositions Across Multiple Tasks.

Skill Ablation We further investigated the direct use of pretrained VLM to generate end-effector waypoints and binary gripper commands by leveraging a knowledge base, manual task descriptions, and background images, with the goal of reducing dependence on atomic skills. As shown in Tab. 3, this approach exhibits relatively low success rates, mainly due to the inherent instability of the VLMs and their inability to effectively handle physical interactions and collisions between the robotic arm and objects. This limitation is particularly evident in complex tasks; for example, the CP task requires large-angle reorientation after opening a drawer, which poses significant challenges for VLM-based waypoint planning. Moreover, this method fails to adequately incorporate contact force information into the manipulation process, rendering it incapable of achieving gentle interactions.

Force-Aware Component Ablation To validate the effectiveness of incorporating force-aware rewards for gentle manipulation, we conduct ablation experiment comparing policies trained with and without explicit force-related penalties in the reward function. During long-horizon executions, contact forces from bilateral tactile sensors were continuously measured, excluding non-contact periods. Fig. 5 illustrates the distribution of contact forces during long-horizon manipulation tasks executed by two different policies. The results indicate that the policy trained with force-aware rewards consistently maintains lower and more stable contact

forces, thereby validating that explicitly incorporating contact force factors effectively promotes gentle manipulation.

Method	OS	OP	CP	PB
Success Rate				
VASK+VT-DP (ours)	0.90	0.73	0.63	0.90
VLM Planning	0.63	0.50	0.00	0.37
Human data+VT-DP	0.30	0.47	0.27	0.60
Average Success Path Length				
VASK+VT-DP (ours)	85	184	179	175
VLM Planning	170	306	/	345
Human data+VT-DP	304	648	763	582
Average Contact Force (N)				
VASK+VT-DP (ours)	0.322	0.170	0.489	0.091
VLM Planning	0.672	0.482	/	0.205
Human data+VT-DP	0.809	0.493	0.932	0.196

Table 3: Simulation Benchmarking Against Extended Baselines: Comparative Evaluation of Success Rate (SR), Average Success Path Length (SPL), and Average Contact Force (ACF) Across Our Method, VLM Planning, and Human Demonstration-Based Approaches.

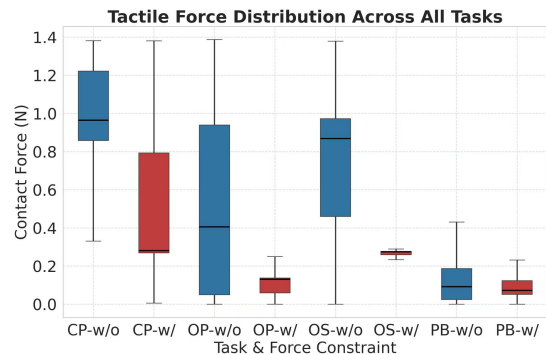


Figure 5: Contact force distribution over ten trials for each of the four manipulation tasks. Blue boxplots show policies trained without force-related penalties, while red boxplots show policies trained with force-aware rewards.

Method Comparisons

Demonstration Data Comparison In simulation, we collected expert demonstrations by manually operating the robotic arm to complete long-horizon tasks and then distilled the data using the VT-DP framework. However, models trained on such demonstrations exhibited suboptimal performance as shown in Tab. 3, often stalling and failing to achieve stable grasping. We attribute this to the low quality of manual data: human demonstrations introduce pauses, inconsistencies, and subjective biases, leading to unnecessarily long and noisy trajectories. As a result, the distilled policies require longer execution paths and show reduced robustness. Manual data collection also incurs significant labor effort and does not effectively encode contact force in-

Method	OS(SR)	OP(SR)	CP(SR)	PB(SR)	Freq(Hz)	Avg(SR)	Task Var
PC(Vision w/ Tactile)	0.90	0.73	0.63	0.90	28	0.79	0.0134
PC(Vision w/o Tactile)	0.83	0.30	0.33	0.86	28	0.58	0.0705
RGB(Vision w/ Tactile)	0.50	0.20	0.00	0.43	9	0.28	0.0389
RGB(Vision w/o Tactile)	0.43	0.23	0.00	0.37	8	0.26	0.0274

Table 4: Success rates of diffusion-based manipulation policies using different perception modalities as observations. For fairness, all methods utilize the same number of expert demonstrations and identical data generation models.

formation. In contrast, our framework efficiently generates smooth, force-aware long-horizon manipulation data with substantially lower collection cost, achieving consistently higher success rates across diverse tasks. These results highlight both the practicality and effectiveness of our approach.

Input Modalities Comparison Robot perception involves both RGB images and point clouds, each offering distinct representational advantages. We systematically evaluated diffusion policies trained on these modalities, conducting 30 trials per task, with all demonstration data generated through the VASK framework to ensure fairness. As shown in Tab. 4, the diffusion policy trained solely on RGB images, represented by the method DP (Chi et al. 2023b), performed poorly across all tasks. Incorporating tactile information brought only marginal improvements, which may be attributed to the inherent difficulty of extracting robust features from RGB data alone, especially in long-horizon manipulation tasks that require precise spatial reasoning. The DP3 method, which utilizes visual point clouds, achieved better results owing to their compact encoding of three-dimensional spatial information. However, it still underperformed on complex, contact-rich tasks such as OP and CP, indicating that spatial data alone cannot fully capture the nuanced tactile interactions these tasks demand. Our proposed approach, VT-DP, which fuses visual and tactile point clouds, consistently outperformed all other methods across tasks. This fusion enhances spatial perception through tactile feedback and improves feature representation, resulting in superior overall task performance.

Qualitative Experiments in the Real World

We deploy our policy on physical robots using a digital twin system, validating the feasibility of our approach. Specifically, policy observations include proprioceptive data and visuo-tactile point clouds. We apply linear transformations to real-world proprioceptive measurements to align with simulation observations. For visual point clouds, we capture real scenes using D415 cameras; however, these point clouds exhibit quality limitations, particularly in robotic arm regions. To address this, we first transform simulated and real-world point clouds into a unified coordinate system, then merge simulated robotic arm point clouds with real-world scene point clouds. The digital twin ensures state consistency between physical and simulated robotic arms. Additionally, we employ color-based thresholding and point cloud cropping to remove irrelevant environmental backgrounds, achieving representations similar to simulation environments. For tactile point clouds, we align GelSight Mini

sensor data with simulated tactile point clouds and perform corresponding downsampling. As shown in Fig. 6, this method enables deployment of our manipulation policies in real-world environments.

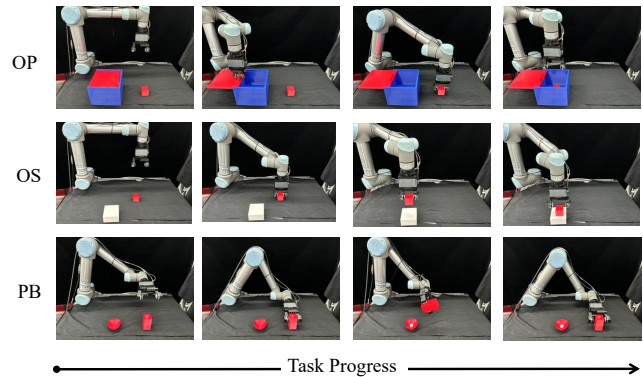


Figure 6: The execution process of long-horizon manipulation tasks in the real world.

Limitations

Although the current framework significantly reduces the manual effort required to generate demonstration data, the integration of VLM incurs additional computational overhead. Future research will focus on narrowing the sim-to-real gap to facilitate effective deployment in real-world scenarios.

Conclusion

In this work, we propose a novel framework that decomposes complex manipulation tasks into compositions of atomic skills and leverages visual language models for task planning and decision-making. Combined with a Visual-Tactile Diffusion Policy (VT-DP), the method successfully distills long-horizon gentle manipulation strategies from expert demonstrations. Notably, our approach explicitly incorporates tactile perception into the generation of long-horizon manipulation data, enhancing the capability of the system to perform delicate and contact-rich operations. Since the policy is fully trained in simulation without requiring additional manual demonstration data, the framework significantly reduces both labor and data collection costs. Additionally, the integration of visual language models with atomic skills provides strong scalability, enabling the framework to generalize across a wide range of long-horizon manipulation tasks.

Acknowledgments

This paper is partially supported by the National Natural Science Foundation of China, grant No.62573159, Guangdong Provincial Key Laboratory of Intelligent Morphing Mechanisms and Adaptive Robotics under Grant 2023B1212010005 and in part by Shenzhen Science and Technology Program under Grant KJZD20240903100802004 and GXWD 20231130153844002.

References

- Akinola, I.; Xu, J.; Carius, J.; Fox, D.; and Narang, Y. S. 2024. TacSL: A Library for Visuotactile Sensor Simulation and Learning. *IEEE Transactions on Robotics*, 41: 2645–2661.
- Ankile, L. L.; Simeonov, A.; Shenfeld, I.; Torné, M. M. L.; and Agrawal, P. 2024. From Imitation to Refinement – Residual RL for Precise Assembly.
- Chebatar, Y.; Vuong, Q.; Hausman, K.; Xia, F.; Lu, Y.; Irpan, A.; Kumar, A.; Yu, T.; Herzog, A.; Pertsch, K.; Gopalakrishnan, K.; Ibarz, J.; Nachum, O.; Sontakke, S. A.; Salazar, G.; Tran, H. T.; Peralta, J.; Tan, C.; Manjunath, D.; Singh, J.; Zitkovich, B.; Jackson, T.; Rao, K.; Finn, C.; and Levine, S. 2023. Q-Transformer: Scalable Offline Reinforcement Learning via Autoregressive Q-Functions. In *7th Annual Conference on Robot Learning*.
- Chen, K.-P.; Xie, S.; Ma, Z.; and Goldberg, K. 2025. Robo2VLM: Visual Question Answering from Large-Scale In-the-Wild Robot Manipulation Datasets.
- Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; and Song, S. 2023a. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems*.
- Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E. A.; Burchfiel, B.; and Song, S. 2023b. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *ArXiv*, abs/2303.04137.
- Geng, H.; and Liu, Y. 2023. UniDexGrasp++: Improving Dexterous Grasping Policy Learning via Geometry-aware Curriculum and Iterative Generalist-Specialist Learning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3868–3879.
- Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1861–1870. PMLR.
- Hua, P.; Liu, M.; Macaluso, A.; Lin, Y.; Zhang, W.; Xu, H.; and Wang, L. 2024. GenSim2: Scaling Robot Data Generation with Multi-modal and Reasoning LLMs. In *Conference on Robot Learning*.
- Huang, B.; Wang, Y.; Yang, X.; Luo, Y.; and Li, Y. 2024. 3D ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing. In *Proceedings of Robotics: Conference on Robot Learning (CoRL)*.
- Jiang, Y.; Yu, M.; Zhu, X.; Tomizuka, M.; and Li, X. 2024. Contact-Implicit Model Predictive Control for Dexterous In-hand Manipulation: A Long-Horizon and Robust Approach. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5260–5266.
- Jin, W.; and Posa, M. 2022. Task-Driven Hybrid Model Reduction for Dexterous Manipulation. *IEEE Transactions on Robotics*, 40: 1774–1794.
- Jing, Z.; Yang, S.; Ao, J.; Xiao, T.; Jiang, Y.; and Bai, C. 2025. HumanoidGen: Data Generation for Bimanual Dexterous Manipulation via LLM Reasoning. *arXiv preprint arXiv:2507.00833*.
- Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanke, P. R.; Vuong, Q.; Kollar, T.; Burchfiel, B.; Tedrake, R.; Sadigh, D.; Levine, S.; Liang, P.; and Finn, C. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *ArXiv*, abs/2406.09246.
- Li, Z.; Niu, Y.; Su, Y.; Liu, H.; and Jiao, Z. 2024. Dynamic Planning for Sequential Whole-body Mobile Manipulation. *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*, 1–7.
- Liu, Z.; Liu, X.; Zhang, Y.; Liu, Z.; and Huang, P. 2023. Tactile Active Inference Reinforcement Learning for Efficient Robotic Manipulation Skill Acquisition. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10884–10889.
- Shi, H.; Xu, H.; Clarke, S.; Li, Y.; and Wu, J. 2023. RoboCook: Long-Horizon Elasto-Plastic Object Manipulation with Diverse Tools. *arXiv preprint arXiv:2306.14447*.
- Wang, L.; Ling, Y.; Yuan, Z.; Shridhar, M.; Bao, C.; Qin, Y.; Wang, B.; Xu, H.; and Wang, X. 2023a. GenSim: Generating Robotic Simulation Tasks via Large Language Models. *ArXiv*, abs/2310.01361.
- Wang, Y.; Xian, Z.; Chen, F.; Wang, T.-H.; Wang, Y.; Erickson, Z.; Held, D.; and Gan, C. 2023b. RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation. *ArXiv*, abs/2311.01455.
- Welle, M. C.; Lippi, M.; Lu, H.; Lundell, J.; Gasparri, A.; and Kragic, D. 2023. Enabling Robot Manipulation of Soft and Rigid Objects with Vision-based Tactile Sensors. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, 1–7.
- Wu, Q.; Peng, X.; Zhou, J.; Sun, Z.; Xiong, X.; and Lou, Y. 2024. RTTF: Rapid Tactile Transfer Framework for Contact-Rich Manipulation Tasks. In *IROS*, 2913–2920.
- Xue, H.; Ren, J.; Chen, W.; Zhang, G.; Fang, Y.; Gu, G.; Xu, H.; and Lu, C. 2025a. Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*.
- Xue, H.; Ren, J.; Chen, W.; Zhang, G.; Fang, Y.; Gu, G.; Xu, H.; and Lu, C. 2025b. Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation. *ArXiv*, abs/2503.02881.

Yang, M.; Lin, Y.; Church, A.; Lloyd, J.; Zhang, D.; Barton, D. A.; and Lepora, N. F. 2023. Sim-to-Real Model-Based and Model-Free Deep Reinforcement Learning for Tactile Pushing. *IEEE Robotics and Automation Letters*, 8(9): 5480–5487.

Yang, Y.; Sun, J.; Kou, S.; Wang, Y.; and Deng, Z. 2025. LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks.

Ze, Y.; Zhang, G.; Zhang, K.; Hu, C.; Wang, M.; and Xu, H. 2024. 3D Diffusion Policy. *ArXiv*, abs/2403.03954.

Zhao, T.; Kumar, V.; Levine, S.; and Finn, C. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. *ArXiv*, abs/2304.13705.

Zhong, Y.; Jiang, Q.; Yu, J.; and Ma, Y. 2025. DexGrasp Anything: Towards Universal Robotic Dexterous Grasping with Physics Awareness. *ArXiv*, abs/2503.08257.