

MaskVLA: Visual Masking Against Trajectory Overfitting of Vision-Language-Action Model

Yuxuan Jiang^{1*}, Jiaying Huang^{1,2*}, Ge Wang^{1,2,3*}, Shenhao Yan^{1,3}, Jiahao Yang^{1,3,4}, Chengsi Yao^{1,3,5}, Qi Liu^{1,3}, Qing Zhao^{1,3}, Shuguang Cui^{2,1}, Yiming Zhao³, Yatong Han^{3†}, Zhen Li^{2,1†}

¹FNii-Shenzhen, The Chinese University of Hong Kong, Shenzhen, China

²School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

³Ising AI ⁴University of Glasgow ⁵School of Automation, Southeast University

Abstract—Vision-Language-Action (VLA) models integrate vision-language understanding with executable robot actions, enabling end-to-end learning for robot control. However, our empirical analysis reveals that existing models exhibit severe trajectory overfitting when finetuned on limited datasets. To guide the model in effectively utilizing wrist camera information, we propose MaskVLA, a masking-based fine-tuning strategy. By randomly masking a small portion of the main camera’s visual information, the model is guided to autonomously learn more fine-grained, task-relevant, and effective visual features. This process leads to the emergence of robust policies, thereby enhancing the model’s capability to tackle complex manipulation tasks and improving its generalization performance. Our method has been comprehensively evaluated on RoboTwin 2.0, achieving an average success rate improvement of 23.2% and 16.8% compared to π_0 and OpenVLA-OFT, respectively. Furthermore, experiments on real-world ALOHA robots also demonstrate the effectiveness of our approach.

I. INTRODUCTION

VLA models have advanced rapidly, demonstrating immense potential in bridging high-level and visual information with precise robotic control. Their ability to generalize across novel scenarios highlights a transformative direction for embodied intelligence. However, VLA models suffer from severe overfitting, memorizing spurious dataset correlations rather than learning robust task semantics.

The model’s brittleness arises from two key factors. First, limited dataset scale and diversity forces reliance on extensive fine-tuning of small datasets to attain acceptable performance. Second, architectural limitations exacerbate domain shift: for instance, OpenVLA-OFT [1] achieves strong performance, yet suffers from severe distribution shift due to mismatch between its single-view pretraining and multi-view fine-tuning, resulting in primary-camera overfitting and trajectory memorization.

We conducted an interpretability analysis using Grad-CAM [2] (Fig. 3) and confirmed that OpenVLA-OFT [1] suffers from a severe overfitting problem: its attention is excessively concentrated on local texture features while largely ignoring multi-view information. This finding demonstrates the model’s overreliance on superficial statistical cues rather than learning the underlying task semantics, revealing a fundamental flaw in its representation learning. Although

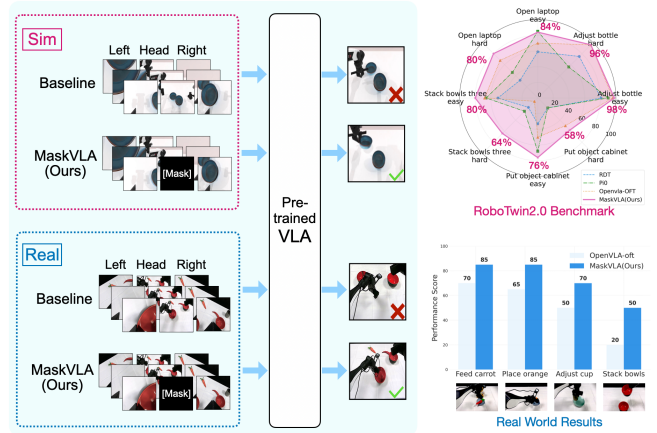


Fig. 1. We propose **MaskVLA**, an improved fine-tuning method for multi-view VLA frameworks that significantly enhances model performance. By randomly masking the primary camera input, the method guides the model to learn robust multi-view visual features. In both RoboTwin2.0 [3] simulation environments and real-world experiments, it demonstrates substantial improvements over OpenVLA-OFT [1] across multiple manipulation tasks.

most VLA training pipelines already incorporate standard visual data augmentation techniques (e.g., random cropping, rotation, blur, and color jitter; see Table II), these augmentations alone are insufficient to resolve the problem. In practice, VLA models still struggle to build a stable global understanding of the scene. Instead, they often overfit to demonstration trajectories, relying on memorized spatial patterns rather than reasoning about object relationships and task goals. As a result, policies frequently exhibit brittle behaviors during execution, such as grasping at incorrect locations (e.g., closing the gripper near but not on the target object), failing to re-localize objects after small perturbations, and lacking the ability to recover from intermediate failures.

Multi-view VLAs require coordinated use of main and wrist cameras for complex tasks. While the main camera provides global context, the wrist camera becomes essential when handling close-range occlusions. Current methods, however, fail to dynamically integrate these views. Instead of predefined switching rules, the model should autonomously learn when and how to use each camera.

Inspired by the effective masking mechanisms in VLMs and LLMs, we propose MaskVLA, a novel fine-tuning method designed for multi-view VLA models. During fine-

*Equal contribution. †Corresponding authors.

Zhen Li: lizhen@cuhk.edu.cn

Project Page: MaskVLA

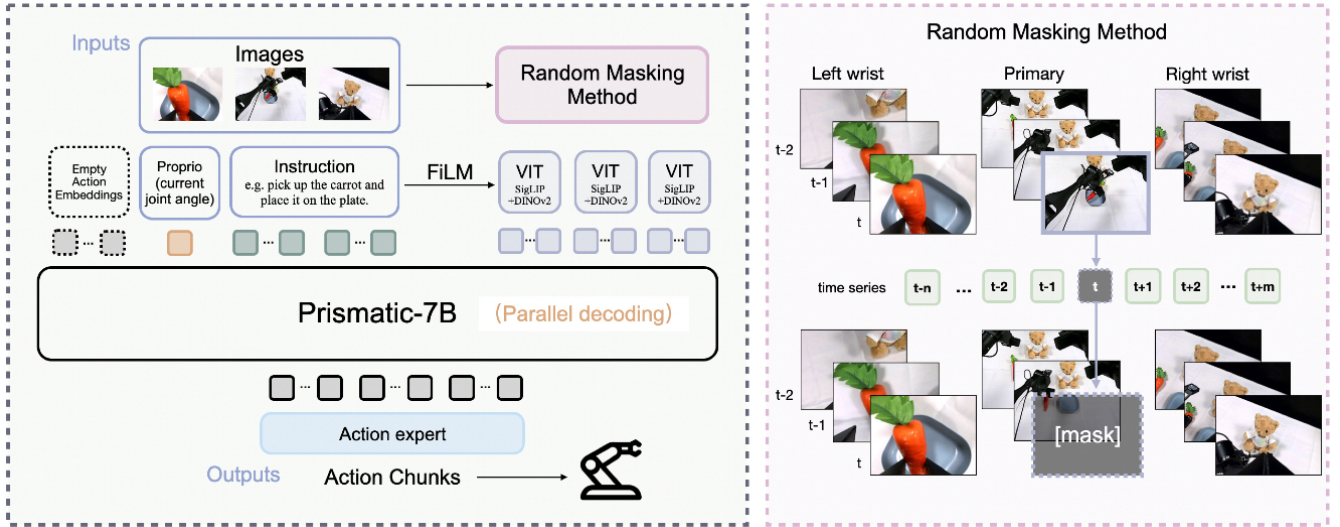


Fig. 2. **Architecture of our MaskVLA.** During the fine-tuning stage, a small portion of the input from the primary camera (e.g., 10%) is masked out. During inference, the model receives complete multi-camera inputs. This masking strategy reduces the model’s dependency on the primary camera and guides more robust and generalized feature learning.

tuning, it randomly masks a small portion of the primary camera input, while during inference, it takes the complete multi-view images as input. This approach encourages the model to focus more on wrist camera information, thereby learning more view-balanced and robust visual features.

Extensive experiments in both simulation and real-world environments validate the efficacy of our method (Fig. 1). In RoboTwin2.0 benchmark [3], our approach significantly outperforms state-of-the-art methods such as OpenVLA-OFT [1] and π_0 [4]. For real-world validation, we deploy our method on an ALOHA dual-arm robot system [5], and the experimental results consistently demonstrate its superior effectiveness and practical applicability.

In summary, we make three major contributions:

- We identify the critical issue of overfitting in SOTA VLA fine-tuning methods, using Grad-CAM to visually confirm the phenomenon.
- We propose MaskVLA, a novel fine-tuning framework to leverage multi-view visual cues, enabling more robust representation learning and precise action generation.
- Experiments in both simulation and real-world confirm the method’s effectiveness in mitigating the problem.

II. RELATED WORKS

A. Visuomotor Policy Learning

The foundation of end-to-end robot control is built upon visuomotor policy learning methods, which map high-dimensional visual observations to low-level robot actions. A wide range of architectures have been proposed, including ACT [5], which employs a Transformer to predict future action chunks; Diffusion Policy (DP) [6], which leverages diffusion models to generate robust action sequences; and its 3D extension DP3 [7], which operates directly on 3D point clouds or voxels. While highly effective in continuous control tasks, these methods typically focus on imitation

learning from demonstrations without incorporating natural language instructions as a conditioning signal, which limits their ability to follow open-ended semantic commands and adapt to tasks requiring linguistic understanding.

B. Vision-language-action Models

Vision-Language-Action (VLA) models [4, 8–12] bridge vision-language understanding with robotic control. Built upon large pre-trained vision-language models (VLMs), they directly map multimodal inputs—visual observations and language instructions—into executable, low-level robot actions. This enables end-to-end learning of language-conditioned visuomotor policies, promising robust generalization to novel objects, environments, and semantic instructions. Representative instances include RDT [10], which employs vision and text encoders to extract multimodal features and utilizes a DiT architecture to model action distributions in latent space, achieving end-to-end generation from perception to action; OpenVLA [9], an open-source VLA foundation model that builds upon pre-trained vision-language models and demonstrates strong performance and generalization across a variety of real-world robotic tasks through its straightforward Transformer-based architecture; and π_0 [4], a foundational VLA that adopts a VLM pre-training plus Flow Matching architecture, demonstrating remarkable cross-embodiment generalization capabilities.

However, existing works largely overlook a critical issue: VLAs fine-tuned extensively on limited datasets for tens of thousands of steps often exhibit severe trajectory memorization, a clear indicator of significant overfitting. Addressing this gap, our work not only identifies this critical overfitting pitfall but also proposes a novel approach to mitigate it, paving the way for more robust VLAs.

C. VLA Fine-Tuning method

Recent advances in VLA fine-tuning [1, 13, 14] focus on improving adaptation efficiency. Zhang et al. [13] systematically study the fine-tuning strategies for Octo [15], identifying key design choices for optimal low-data performance. ReVLA [14] reconstructs SigLIP [16] and DINOv2 [17] weights to improve VLA out-of-distribution robustness. OpenVLA-OFT [1] achieves state-of-the-art results by efficiently transforming single-arm models into bimanual multi-view policies. However, it suffers from overfitting and poor generalization due to a mismatch between single-camera pretraining data and triple-camera fine-tuning, leading to trajectory memorization. Our approach effectively mitigates this issue.

D. Masking in LLM/VLM

Masking is a core self-supervised learning method in LLMs and VLMs. In text domains, BERT [18] significantly enhances language models’ semantic understanding through bidirectional context prediction. In visual domains, MAE’s [19] high masking ratio (e.g., 75.0%) and block-wise masking design compel the model to learn global semantic reasoning rather than local texture replication, markedly improving the generalization capability of visual representations. Recent studies (e.g., MaskGit [20]) have further introduced bidirectional masked prediction mechanisms into image generation, employing non-autoregressive mask-iterative reconstruction strategies to maintain generation quality while substantially increasing inference speed. These advances indicate that masked modeling is evolving into a unified paradigm bridging understanding and generation tasks.

Inspired by the approaches, we introduce this method into the fine-tuning of VLA models and propose **MaskVLA**, which enhances both the generalization capability and accuracy of the model.

III. METHODOLOGY

We propose an improved fine-tuning framework for multi-camera Vision-Language-Action models (VLAs). Motivated by empirical evidence of overfitting in standard fine-tuning, our core contribution is a plug-and-play Random Masking on Primary Camera View module for multi-view VLA finetune, called **MaskVLA**. It effectively regularizes the learning process, leading to significant gains on challenging downstream tasks.

A. Overfitting Problem in VLA

In generalist robot policies trained on large-scale, multi-source datasets, we observe a pronounced overfitting phenomenon. This overfitting refers to the policy’s tendency to memorize spurious statistical correlations – those idiosyncratic to the composition of the training dataset (e.g., specific background textures, lighting conditions, or robot kinematic properties) – rather than learning the underlying, invariant task semantics. Consequently, despite being exposed to vast amounts of data, the model remains brittle to any distribution shift that disrupts these memorized correlations.

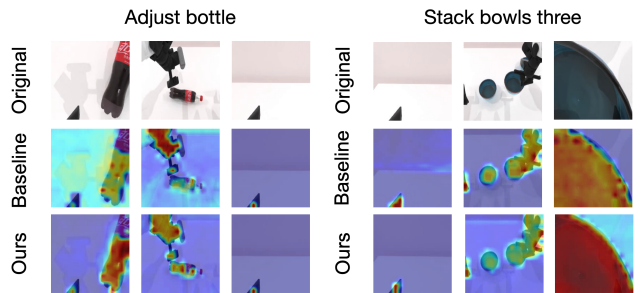


Fig. 3. **Grad-CAM Analysis:** Qualitative comparisons on two RoboTwins tasks. **Row 1:** Input images. **Row 2:** Baseline (OpenVLA-OFT) shows scattered attention on irrelevant areas. **Row 3:** MaskVLA concentrates attention precisely on task-critical regions validating its superior visual grounding via masked training.

Our analysis reveals this overfitting originates from two sources: (1) fundamental dataset limitations including insufficient diversity and fragmentation [21], which promotes learning of narrow sub-distribution features; and (2) a critical architectural deficiency in OpenVLA-OFT [1] where the model—pre-trained on single-arm, single-view data—is directly fine-tuned for dual-arm, multi-view tasks without commensurate pre-training. This creates a severe domain gap, forcing the policy to default to its pre-training bias and exhibit over-reliance on the primary camera view rather than developing balanced attention across all visual inputs.

To empirically validate the observed overfitting phenomenon, we conducted an interpretability analysis of the attention distribution in the model’s visual encoder. Specifically, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) [2] to visualize the fine-tuned OpenVLA-OFT [1] model.

As shown in Fig. 3, the heatmaps clearly reveal a pronounced visual preference bias: the model’s attention is highly concentrated on local texture features within the primary camera view (e.g., logos on the robot arm surface, workstation edges), while largely ignoring information from multiple auxiliary cameras. This imbalanced attention distribution confirms that the model indeed suffers from overfitting dependence on the primary view. It has learned to leverage superficial statistical features specific to this perspective rather than understanding the underlying physical principles of the task.

This finding provides computational visual evidence supporting our theoretical analysis, indicating that, under the current experimental setup, primary-view dependence primarily arises from limitations in representation learning, rather than being solely due to data volume.

To address the aforementioned over-reliance on the primary camera, we propose a novel and effective solution: Random Masking on Primary Camera View (MaskVLA). This technique is designed to explicitly discourage the policy from fixating on any single viewpoint during training. By randomly masking visual observations from the primary camera with a predetermined probability, we force the policy to learn to integrate information from all available visual streams dynamically and robustly. This approach effectively

mitigates the pre-training bias and closes the domain gap by ensuring that the model does not become dependent on a potentially privileged sensory input.

B. MaskVLA: Random Masking on Primary Camera View

Our method builds upon the OpenVLA-OFT [1] framework, a state-of-the-art approach for finetuning Vision-Language-Action policies. OpenVLA-OFT employs parallel decoding, continuous action prediction with an L1 regression objective, and FiLM [22] modules to efficiently process multi-view visual inputs.

At each timestep t , the robot receives a multimodal observation defined as

$$O_t = \{P_t, I_p^t, I_l^t, I_r^t\}, \quad (1)$$

where P_t represents the robot’s proprioceptive state, specifically the 14-dimensional joint angles, and I_p^t, I_l^t, I_r^t correspond to RGB images from the primary (overhead) camera and the left/right wrist cameras. The primary view provides a global overview of the scene, while the wrist views offer detailed local information that is essential for precise manipulation.

Based on these observations, the VLA policy π_θ predicts a sequence of future actions over a horizon H , conditioned on the current observation and task instruction $\hat{\ell}$:

$$\pi_\theta(a_{t:t+H} \mid O_t, \hat{\ell}), \quad (2)$$

where $a_{t:t+H}$ denotes the predicted action sequence. This formulation allows the policy to leverage multi-view visual inputs, proprioceptive feedback, and high-level language guidance to generate coordinated manipulation behaviors.

Despite this multi-view design, existing VLA policies often develop a strong bias toward the primary camera during training, leading to poor generalization when the global view changes or becomes occluded. To mitigate this issue, we introduce **MaskVLA**, a simple stochastic masking strategy applied to the primary camera during training.

Specifically, at each discrete timestep t_i , the primary view is either fully masked or left intact according to a Bernoulli random variable:

$$I_p^{t_i, \text{input}} = (1 - M_i) \odot I_p^{t_i}, \quad M_i \sim \text{Bernoulli}(\rho), \quad (3)$$

where $M_i = 1$ indicates that the primary view is masked and ρ denotes the masking probability. The masking operation $\mathcal{M}(\cdot)$ sets all pixel values of the primary image to zero, producing the masked image $I_p^{t_i, [\text{mask}]}$.

During training, the masked primary image $I_p^{t_i, [\text{mask}]}$ is processed together with the left- and right-wrist images $I_l^{t_i}$ and $I_r^{t_i}$ by the visual encoder to extract viewpoint-specific features F_p, F_l , and F_r . These features are then concatenated along the sequence-length dimension to form a fused visual representation F , which is fed into the VLA policy network π_θ . The resulting observation used for policy learning is therefore

$$O_t = \{I_p^{t_i, [\text{mask}]}, I_l^{t_i}, I_r^{t_i}, P_t\}. \quad (4)$$

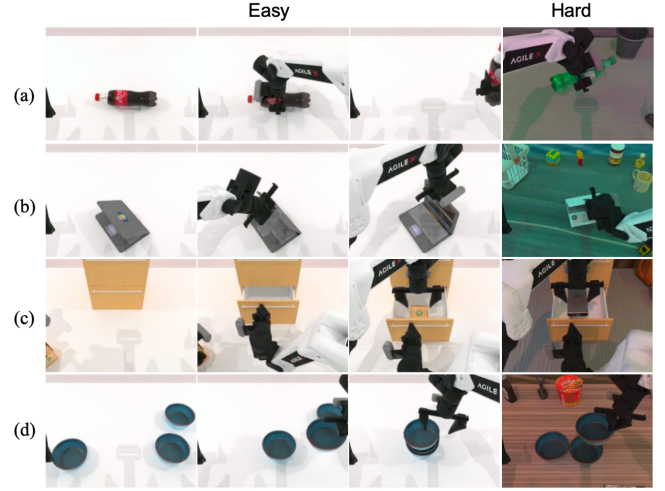


Fig. 4. **Representative tasks from the RoboTwin2.0 benchmark.** Each row corresponds to a task: (a) Adjust bottle, (b) Open laptop, (c) Put object cabinet, and (d) Stack bowls three. The first three columns illustrate task progression in the easy setting (initial, interaction, goal), while the rightmost column shows a representative state in the hard setting with randomized object poses and layouts, evaluating policy generalization under out-of-distribution conditions.

By randomly masking the primary view during training, the model is encouraged to utilize wrist-view information more effectively, which is often crucial for fine-grained manipulation or situations involving occlusion. This strategy simulates scenarios where the global view is unavailable and forces the policy to learn more balanced multi-view representations. During inference, the masking operation is disabled, allowing the policy to exploit the full information from all cameras and achieve optimal performance.

IV. SIMULATED EXPERIMENT IN ROBOTWIN

Our Simulated experiments investigated the following questions:

- How does the masking method improve model performance?
- Does the masking method help enhance the generalization ability of models?
- What is the optimal masking strategy?

A. Experimental Setup

1) *Simulation Environment:* We primarily conduct experiments on the **RoboTwin2.0** [3] simulation benchmark, using the ALOHA-AgileX dual-arm system as the manipulator to evaluate our model’s effectiveness in bimanual collaboration tasks. RoboTwin provides an excellent simulation validation environment for the proposed method. It offers a complete pipeline for data generation, algorithm verification, and evaluation.

RoboTwin provides two environmental modes, easy and hard. Easy mode (demo_clean) is in a white-background, clutter-free environment and primarily tests model task execution capability, serving as an upper bound for task execution accuracy. Hard mode (demo_randomized) tests model capability by approximating the rich variability of real environments through randomly generating desktop and

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON ROBOTWIN BENCHMARK

Method \ Success Rate(%)	Adjust bottle		Open laptop		Put object cabinet		Stack bowls three		Average $\uparrow$		
	easy	hard	easy	hard	easy	hard	easy	hard	easy	hard	all
ACT [5]	27.0	14.0	32.0	0.0	0.0	0.0	0.0	0.0	14.8	3.5	9.1
DP [6]	97.0	0.0	49.0	0.0	42.0	0.0	63.0	0.0	62.8	0.0	31.4
DP3 [7]	99.0	3.0	82.0	7.0	<u>72.0</u>	1.0	57.0	5.0	<u>77.5</u>	4.0	40.8
RDT [10]	81.0	<u>75.0</u>	59.0	32.0	33.0	18.0	51.0	17.0	56.0	35.5	45.8
π_0 [4]	90.0	56.0	85.0	46.0	68.0	18.0	66.0	<u>24.0</u>	77.3	36.0	56.6
OpenVLA-OFT [1]	96.0	96.0	70.0	<u>68.0</u>	50.0	<u>50.0</u>	<u>68.0</u>	6.0	71.0	<u>55.0</u>	<u>63.0</u>
MaskVLA(Ours)	<u>98.0</u>	96.0	<u>84.0</u>	80.0	76.0	58.0	80.0	66.0	84.5 $\uparrow$ 7.0	75.0 $\uparrow$ 20.0	79.8 $\uparrow$ 16.8

background textures, randomly generating lighting, serving as a more realistic evaluation environment.

2) *Task Description*: We selected four tasks require information from both the main camera and wrist cameras to complete. As shown in the pictures below Table I, the tasks are as follows:

Adjust bottle: requires the robot to adjust the position and orientation of a bottle. Base on the bottle’s initial random orientation, the model needs to determine whether to use the left or right hand for grasping.

Open laptop: requires necessitating precise control of the opening angle, testing the model’s ability to learn curved motion trajectories with obstructing the main camera’s view.

Put object cabinet: involves a complex sequence of actions including opening cabinet, grasping, and placing. Different sub-steps have varying degrees of dependence on the main camera and wrist cameras respectively, testing the model’s attention capability across the three cameras.

Stack bowls three: tests the model’s generalization capability for long-horizon manipulation task. Unlike Put Object Cabinet, this task operates on identical objects.

3) *Training Protocol*: All policies are trained on the Aloha-AgileX embodiment using 50 `demo_clean` trajectories per task. Specifically, all pre-trained policies (e.g., π_0 and OpenVLA-OFT) are fine-tuned on each single task based on their released open-source weights. Evaluation is conducted over 100 trials under both the `demo_clean` (Easy) and `demo_randomized` (Hard) settings. Unless otherwise specified, all baseline methods adopt their default data augmentation strategies as provided in the original implementations. Details of the data augmentation strategies adopted by different baseline methods are summarized in Table II.

B. Experimental Results Analysis

1) *Evaluation Results Analysis*: Among the selected 8 sub-tasks, the MaskVLA method achieved the best performance in 6 sub-tasks and obtained second-best performance in the remaining 2 tasks. In particular, our method achieved the highest average success rate in both Easy and Hard modes across 4 tasks, fully validating the effective improvement of the masking strategy on model performance.

In Hard mode, scene lighting and distractors vary randomly, further testing the model’s scene generalization capa-

TABLE II
COMPARISON OF DATA AUGMENTATION STRATEGIES

Method	Random Crop ¹	Color Jitter ²	Rotate ³	Noise / Blur ⁴	Random Masking ⁵
ACT [5]					
DP [6]	✓				
DP3 [7]					
RDT [10]		✓		✓	
π_0 [4]	✓	✓	✓		
OpenVLA-OFT [1]	✓	✓			
Ours	✓	✓			✓

¹Random crop and resize. ²Randomly alters image brightness, contrast, saturation, and hue. ³Random rotation within a small angle. ⁴Adds random noise or blur. ⁵Randomly masks out primary view image patches (our method).

bility. Experimental results show that MaskVLA achieved an average success rate of 75.0% in Hard mode, outperforming the baseline method OpenVLA-OFT by 20.0%. It confirms that our method can effectively enhance the model’s generalization capability.

From the specific task in Hard mode performance analysis in Table 1, our method demonstrated differentiated improvement effects across different tasks. Among them, the Stack bowls three task showed the most significant improvement, with a success rate increase of 60.0% compared to the OpenVLA-OFT, clearly surpassing the π_0 by 42%. For the Open laptop and Put object cabinet tasks, our method achieved performance gains of 12.0% and 8.0% respectively, outperforming the OpenVLA-OFT. For the Adjust bottle task, where OpenVLA-OFT already performed well, our method still maintained the same high success rate without performance degradation.

The fundamental reason why the Stack bowls three task achieved such significant improvement lies in its high requirements for position generalization capability. Although the Adjust bottle and Open laptop tasks involve random position generation, the operation targets are all single objects; in the Put object cabinet task, the variation range of object grasping positions is relatively limited, with a low degree of position generalization. In contrast, the Stack bowls three task needs to simultaneously handle three randomly distributed operation objects, placing higher demands on the model’s spatial understanding ability and the generalizability of sequential operations. Therefore, our method

TABLE III
ABLATION EXPERIMENT RESULTS (SUCCESS RATES %)

Method	Task				Avg
	Adjust bottle	Open laptop	Stack bowls	Put object	
Baseline [1]	96.0	68.0	6.0	50.0	55.0
MaskVLA-patch	96.0	<u>72.0</u>	30.0	40.0	59.5
MaskVLA-camera	<u>94.0</u>	60.0	58.0	<u>54.0</u>	66.5
MaskVLA-main-5%	96.0	66.0	66.0	50.0	69.5
MaskVLA-main-10%	94.0	80.0	<u>64.0</u>	58.0	<u>74.0</u>
MaskVLA-main-20%	84.0	62.0	14.0	40.0	50.0
Our results	96.0	80.0	66.0	58.0	75.0 $\uparrow 20.0$

Note: **Bold** indicates best performance, underlined indicates second-best. Success rates are reported for all four tasks in hard mode.

demonstrated the most outstanding performance advantage in this most challenging position generalization scenario.

C. Ablation Study

To determine the optimal masking strategy, we designed and implemented systematic ablation experiments. All results reported are evaluated in **hard mode** to rigorously test the model’s generalization to challenging and out-of-distribution scenarios. We proposed three different masking strategies for comparative evaluation:

- 1) **MaskVLA-patch**: During training, the main camera image input is segmented into patch sets after passing through the Vision Transformer, with each patch independently masked with a 10% probability;
- 2) **MaskVLA-camera**: During training, at each time step, one perspective is randomly selected from the main camera view and left/right wrist camera views, and masked with a 10% probability;
- 3) **MaskVLA-main- $\rho\%$** : During training, at each time step, the complete image from the main camera perspective is masked with a preset probability $\rho\%$.

The experimental results in Table III show that all models trained with masking strategies achieved average success rates approaching or exceeding the baseline method’s 55.0%, fully validating the universal effectiveness of masking methods.

Through comparative analysis, we found that the **MaskVLA-main-10%** performed optimally, achieving task success rates above 74.0% under appropriate mask ratio configurations, outperforming **MaskVLA-patch** and **MaskVLA-camera**. Further parameter sensitivity analysis revealed that different tasks have varying optimal mask ratios, suggesting the need for adaptive parameter tuning according to specific task characteristics.

The ablation experimental results confirmed that the **MaskVLA-main- $\rho\%$** is an efficient model optimization strategy. This method selectively occludes main perspective information during training, forcing the model to learn better integration of multi-view features, thereby significantly improving generalization performance in complex manipulation tasks.

Based on the comprehensive simulation experimental analysis above, we provide the following answers to the key questions raised at the beginning of the research:

- By forcing the model to integrate multi-view information during training, the masking strategy effectively addressed the overfitting problem of excessive reliance on the main camera.
- In tasks have high position generalization requirements, MaskVLA achieved the highest average success rate, fully demonstrating the significant enhancement effect of the masking method on model generalization capability.
- Comparative experiments showed that our **main camera masked strategy (MaskVLA-main- $\rho\%$)** is the most effective method, selectively occludes main camera information, encouraging the model to better utilize fine manipulation information provided by wrist cameras, leading strong performance advantage in complex manipulation tasks.

V. EXPERIMENTS IN REAL WORLD

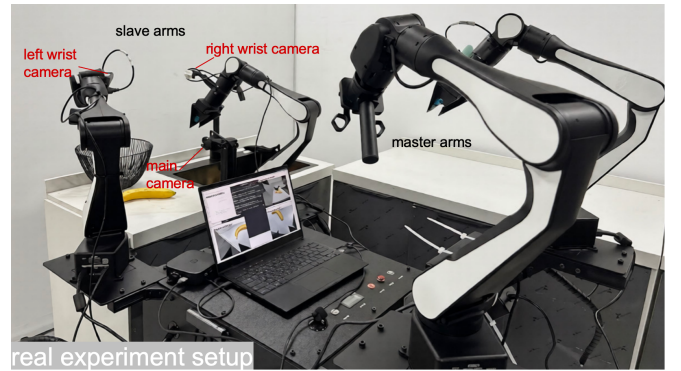


Fig. 5. **Real-world experimental setup.** The platform includes master and slave 6-DoF arms with 1-DoF grippers and three cameras. The master arm is used for teleoperation, while the slave arm executes tasks, enabling multi-view data collection for evaluating VLA policies.

A. Experimental Setup

Experiments are conducted on the Airbot Mobile ALOHA dual-arm platform (Fig. 5). The robot is controlled using

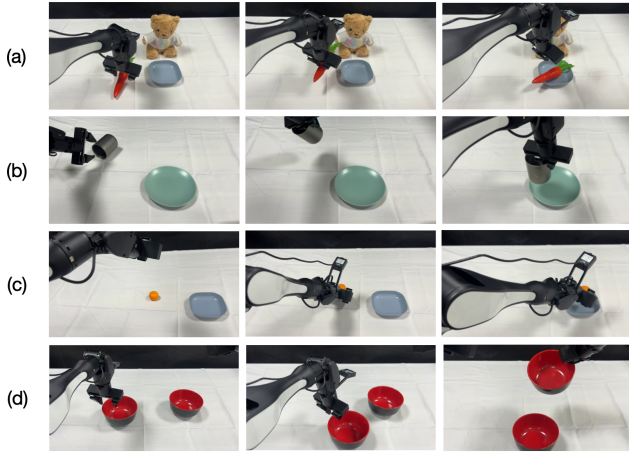


Fig. 6. **ALOHA robot experimental task demonstration.** Four representative tasks test MaskVLA’s multi-view information integration capability: (a) Feed carrot; (b) Adjust cups; (c) Place orange; (d) Stack bowls.

proprioceptive and visual inputs, including 14-dimensional joint angles and images from three cameras.

We collect 50 trajectories per task. For evaluation, each model is tested on each task with 20 trials, and the success rate is used as the performance metric. Training follows the same protocol as in simulation experiments.

B. Task Design

We evaluate MaskVLA on four representative real-world manipulation tasks (Fig. 6) to test multi-view perception and precise control:

Feed carrot: Place a carrot onto the food plate in front of a small bear. The elongated shape requires strong reliance on wrist-camera views during grasping and placement.

Adjust cups: Pick up cups on the desktop, align them upright, and place them at the center of a plate. Occlusions from the robotic arm necessitate effective use of wrist-camera observations.

Place orange: Pick up a small orange and place it in a fruit dish, evaluating precision and dexterity for small objects.

Stack bowls: Stack bowls using both arms, testing spatial understanding and bimanual coordination.

C. Experimental Results Analysis

As shown in Fig. 7, MaskVLA consistently outperforms the state-of-the-art VLA fine-tuning method OpenVLA-OFT across all four real-world tasks. In particular, MaskVLA achieves success rates of 85.0% on both the Feed carrot and Place orange tasks and 70.0% in Adjust cup, outperforming the baseline. While OpenVLA-OFT demonstrates limited effectiveness in executing challenging, long-horizon manipulation tasks (e.g. Stack bowls), our MaskVLA shows substantial performance gains through balanced vision representation learning. Benefiting from the random masking strategy, MaskVLA effectively utilizes both the information from the primary view and wrist view to complete complex tasks and mitigate the overfit problem, highlighting the enhanced visual generalization capability of our approach.

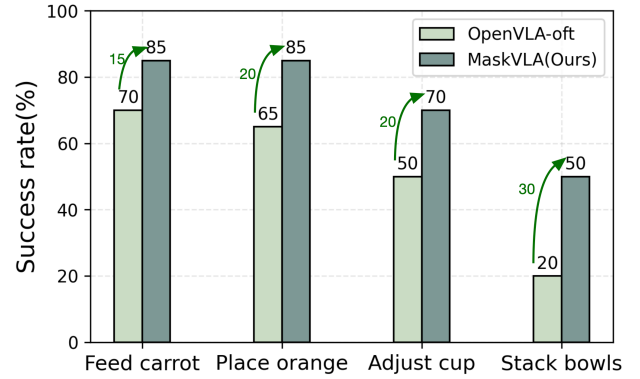


Fig. 7. **Real-world multi-task evaluation comparing the baseline and our method.** The bar chart shows task-wise success rates, highlighting the effectiveness of our approach in improving performance across all tasks under real-world conditions.

VI. CONCLUSIONS

In this paper, we identify and address the overfitting issue in current Vision-Language-Action models (VLAs), which often compromises precise manipulation capabilities. To mitigate this, we propose **MaskVLA**—a novel fine-tuning method that incorporates a random masking strategy applied to the primary view, facilitating more balanced multi-view visual representation learning. Our approach effectively promotes active visual attention allocation and significantly improves manipulation performance. Extensive experiments in both simulated and real-world environments demonstrate the superiority of the proposed method. Furthermore, by evaluating various masking techniques, we find that applying minimal masking to the main camera view yields the best results.

IMPLEMENTATION DETAILS

Models were fine-tuned on 2 NVIDIA A100 GPUs (batch size 4 per GPU) for 30,000 steps per task (~12 hours). Experiments were conducted on RoboTwin2.0 with an RTX 3090 for evaluation and an A100 GPU for inference.

ACKNOWLEDGMENT

This work was supported in part by the Guangdong S&T Programme (Grant No. 2024B0101030002), the Basic Research Project of Hetao Shenzhen-HK ST Cooperation Zone (No. HZQBKCZY-2021067), the Shenzhen Outstanding Talents Training Fund (No. 202002), and the NSFC (Grant Nos. 62293482, 62573371, 61931024, and 12326610). This work was also supported by the Guangdong Province Radio Science Data Center (Grant No. 2025B1212070001), the Shenzhen General Program (No. JCYJ20220530143600001), the Guangdong Research Projects (Nos. 2017ZT07X152 and 2019CX01X104), the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. SYSPG20241211173853027), the National Key Research and Development Program of China (Nos. 2025YFF0515300 and 2025YFF0515304), the Open Project Program of Key Laboratory of Tibetan Information

Processing, Ministry of Education (Grant No. QHSFCS-2606), the Shenzhen-Hong Kong Joint Funding (No. SGDX20211123112401002), and the Tencent & Huawei Open Fund (Nos. 2024E0009 and 202301030019).

REFERENCES

- [1] Moo Jin Kim, Chelsea Finn, and Percy Liang. “Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success”. In: *Robotics: Science and Systems XXI* (2025).
- [2] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017.
- [3] Tianxing Chen et al. *RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation*. arXiv: 2506.18088.
- [4] Kevin Black et al. “ π_0 : A Vision-Language-Action Flow Model for General Robot Control”. In: *Robotics: Science and Systems XXI* (2025).
- [5] Tony Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. “Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware”. In: *Robotics: Science and Systems XIX* (2023).
- [6] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. “Diffusion policy: Visuomotor policy learning via action diffusion”. In: *The International Journal of Robotics Research* (2023).
- [7] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. “3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations”. In: *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*. 2024.
- [8] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. “UniVLA: Learning to Act Anywhere with Task-Centric Latent Actions”. In: *Robotics: Science and Systems XXI* arXiv:2505.06111 (2025). arXiv: 2505.06111.
- [9] Moo Jin Kim et al. “OpenVLA: An Open-Source Vision-Language-Action Model”. In: *Conference on Robot Learning (CoRL), 2024*. PMLR, 2024.
- [10] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. “RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation”. In: *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.
- [11] NVIDIA et al. “GR00T N1: An Open Foundation Model for Generalist Humanoid Robots”. In: arXiv:2503.14734 (). arXiv: 2503.14734.
- [12] Delin Qu et al. “SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model”. In: *Robotics: Science and Systems XXI* arXiv:2501.15830 (2025). arXiv: 2501.15830.
- [13] Wenbo Zhang, Yang Li, Yanyuan Qiao, Siyuan Huang, Jiajun Liu, Feras Dayoub, Xiao Ma, and Lingqiao Liu. “Effective tuning strategies for generalist robot manipulation policies”. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2025.
- [14] Sombit Dey, Jan-Nico Zaeck, Nikolay Nikolov, Luc Van Gool, and Danda Pani Paudel. “Revla: Reverting visual domain limitation of robotic foundation models”. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2025.
- [15] Dibya Ghosh et al. “Octo: An Open-Source Generalist Robot Policy”. In: *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*. Ed. by Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado. 2024.
- [16] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. “Sigmoid Loss for Language Image Pre-Training”. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 11941–11952.
- [17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. “Dinov2: Learning robust visual features without supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).
- [18] Jacob Devlin Ming-Wei Chang Kenton, Lee Kristina Toutanova, et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of naacL-HLT*. Minneapolis, Minnesota. 2019.
- [19] 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*. 2022.
- [20] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. “Maskgit: Masked generative image transformer”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
- [21] Youguang Xing, Xu Luo, Junlin Xie, Lianli Gao, Hengtao Shen, and Jingkuan Song. *Shortcut Learning in Generalist Robot Policies: The Role of Dataset Diversity and Fragmentation*. arXiv: 2508.06426.
- [22] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. “Film: Visual reasoning with a general conditioning layer”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.