HALO: Long Horizon Latent Action Learning for General Robot Manipulation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Robotic manipulation often requires understanding long-horizon tasks guided by visual observations and language instructions. However, most existing Vision-Language-Action (VLA) models focus primarily on short-horizon tasks and overlook the rich historical video context, limiting their ability to perform complex, multi-step tasks. Moreover, these models often suffer from weak alignment between pre-trained vision-language embeddings and robotic actions, which hinders the effective extraction of action-relevant priors from visual input and leads to inaccurate action generation. In this paper, we propose a novel Long **H**orizon Latent **A**ction **L**earning framework for general r**o**bot manipulation, **HALO**, which enables robots to perform multi-step tasks by integrating long-term visual observations, multi-view camera images, and natural language instructions. To capture long-term dependencies, we propose to incorporate Qwen2.5-VL capable of processing long video and multi-view image sequences conditioned on natural language instructions. We further propose the State-Aware Latent Re-representation, which leverages robot states to query action-relevant features by selectively compressing and filtering the vision-language representations. The selected action-aligned embeddings are subsequently fed into an action expert, which predicts multi-step actions via a progressive denoising process. We have trained one of the largest VLA models with 10B trainable parameters, which is first pre-trained on one million diverse real-world robot episodes and fine-tuned across a wide range of downstream tasks. Experimental results on both simulated and real-world tasks demonstrate that our method achieves superior performance compared to prior state-of-the-art methods, particularly in long-horizon manipulation tasks.

1 Introduction

With the rapid development of vision-language models (VLM) [1, 2, 3], robot manipulation policy models have seen significant progress. One of the most active areas in this domain is the Vision-Language-Action (VLA) model, which enables robots to perform complex tasks guided by natural language instructions [4, 5]. Notably, VLA models demonstrate strong generalization capabilities, even in environments that differ from the training distribution. This impressive performance is largely attributed to the powerful cross-modal understanding and reasoning abilities of VLMs, which allow the models to interpret diverse visual scenes and comprehend complex language commands within a unified framework.

The key to successfully training large Vision-Language-Action (VLA) models lies in effectively adapting vision-language models (VLMs) to a wide range of robotic manipulation tasks and designing task-specific components to generate accurate actions. Some approaches [4, 6] fine-tune VLMs to produce discrete action tokens, leveraging their large-scale pretrained knowledge while

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

preserving reasoning capabilities. Although these methods support generalized manipulation skills, the quantization process disrupts the continuity of actions. Other methods [5, 7, 8, 9] introduce a diffusion-based action head on top of the VLM. These models use vision-language embeddings extracted by the VLM as conditional inputs to iteratively denoise probabilistic noise into future actions. However, because vision-language embeddings and actions originate from different modalities, these methods often suffer from weak cross-modal alignment. Directly fusing different modalities may introduce action-irrelevant information, such as background noises or visual distractors, into the decision process, and thus hinders accurate action prediction. Moreover, most existing models rely solely on the current frame to guide the robot, neglecting the importance of historical context. A single frame captures only the present state and overlooks temporal consistency, which can result in discontinuous or unstable actions. Incorporating historical information is essential, as it enables the model to generate more coherent action sequences and enhances its robustness in complex or dynamic environments.

To address these issues, we propose a novel Long **H**orizon Latent **A**ction **L**earning framework for general r**o**bot manipulation, **HALO**, specially designed for long-horizon robotic tasks. **HALO** supports both extended historical frame sequences and multi-view camera inputs, and is capable of extracting action-relevant information from vision-language embeddings to guide future action generation. To fully leverage historical context, we adopt Qwen2.5-VL to process long video sequences and multi-view images. We select Qwen2.5-VL because of its strong ability to understand extended visual contexts, enabling the model to capture rich temporal and spatial information critical for long-horizon manipulation tasks. We further introduce a **Selective Spatial-Temporal Sampling** strategy that effectively integrates historical frames from multiple camera views. Processing all historical frames at high resolution incurs substantial computational cost, and not all frames contribute equally to decision-making. Our strategy is designed to minimize information loss while optimizing memory efficiency. Specifically, we reduce the resolution of historical frames from the primary view while preserving the full resolution of current-frame images across all views. This approach strikes a careful balance between leveraging rich historical context and retaining high-fidelity current observations, ultimately enhancing model performances in complex, long-horizon scenarios.

In addition, studies have shown that robot state information, such as joint angles and end-effector positions, shares the same modality as the action output, making it beneficial for action generation. For example, π_0 leverages both robot state and vision-language embeddings to guide action prediction [5]. However, effectively fusing robot state with visual and language conditions remains challenging. Visual content is often high-dimensional and redundant compared to the compact action modality, which can lead to ineffective fusion and ultimately limit the accuracy of action prediction. To address this issue, we propose a **State-Aware Latent Re-representation** that leverages state information of robots to extract and refine the most action-relevant features from vision-language embeddings, thereby providing more accurate guidance for action generation. Specifically, we first propose a latent space generation method that computes the pairwise product between each token in the state embedding and each token in the vision-language embedding. This results in a large feature space that facilitates the search for action-relevant information. Then, we introduce a learnable mask that suppresses action-irrelevant information while preserving action-relevant cues from the latent space. This process transforms the vision-language embeddings into action-aligned embeddings that exhibit both modality consistency and strong action relevance.

We scale our model up to **10 billion trainable parameters** with optimized training strategies, which significantly enhances its capabilities in both perception and action generation. To enhance generalization, we adopt a step-by-step training pipeline that begins with large-scale pretraining and is followed by task-specific fine-tuning. In the pretraining phase, the model is trained on a large and diverse cross-embodiment robotic dataset comprising one million episodes, combining data from OXE [10] and the AgiBoT dataset [11]. This is followed by fine-tuning on three simulation datasets and real-world data collected using a Franka Research 3 robot setup. This training strategy enables the model to achieve state-of-the-art performance across a wide range of manipulation tasks and demonstrates strong generalization capabilities in handling long-horizon scenarios.

2 Related Work

2.1 Vision-Language-Action Model

Recently, relying on the powerful understanding and reasoning capabilities of Vision-Language-Models (VLM), Vision-Language-Action (VLA) models have made rapid progress, which integrates the action generation for adapting the robot manipulation tasks. For example, RT-2 [12] fine-tunes the VLM on large-scale vision-language data and robotic demonstration data using next-token prediction. It discretizes robotic actions into 256 binary values and represents them as independent tokens similar to text tokens. OpenVLA [4] adopts a similar discretization approach to fine-tune the Prismatic VLM [13] on the Open X-Embodiment dataset [10]. π_0 [5] consists of a PaliGemma model [14] and a separate action expert module, where the VLM is responsible for scene understanding, and the action expert module generates continuous actions through flow matching. Notably, while these models have shown some zero-shot ability, they usually use a single frame and ignore the temporal relationships, which may hinder the models generate consecutive actions and finally result in the failure of the task.

2.2 Diffusion-based Robot Policy

The diffusion model [15, 16, 17] is a mainstream model in the field of image generation. Recent studies [18, 19] have shown that diffusion models can effectively simulate various feasible trajectories that a robot may take to solve a given task. Diffusion policy [18] represents the visuomotor policy of robots as a conditional denoising diffusion process. Inspired by diffusion policies, Octo [20] incorporates a small diffusion head with a 3M parameters into a transformer-based backbone architecture to adapt the action outputs of different robots. RDT [21] proposes a pioneering diffusion foundation model for bimanual manipulation, with the diffusion model reaching 1 billion parameters. CogACT [7] first uses VLM to generate cognition tokens, then uses them as conditions to guide the diffusion model in generating actions that the robot can understand. However, these methods use vision-language embeddings that are not aligned with actions as conditions to guide action generation. In contrast, our model first aligns the vision-language embeddings with the state information of robots and achieves superior results.

2.3 Long-Horizon Robot Manipulation

In the field of robotic manipulation, learning long-horizon tasks has long been a persistent challenge [22, 23, 24, 25, 26, 27]. These tasks typically involve a series of fine-grained actions, each of which must account for physical constraints and their potential consequences, making them highly challenging for the policy model. For example, a long-horizon task may involve opening a microwave, placing a bowl of milk inside, closing the door, and setting the timer for 10 seconds. When task demonstrations are available, many studies, including PerAct [25], ARM [24], and VAPO[28], attempt to decompose complex long-horizon tasks into multiple stages by identifying sub-goals, thereby providing intermediate learning signals and mitigating the accumulation of action errors. However, these decomposition strategies often rely on task-specific knowledge, making them difficult to generalize to new tasks. Besides, ReflectVLM [26] aims to predict future world states and use these predictions to guide action selection and error correction, while DTP [27] attempts to adapt to long-horizon tasks by forecasting the trajectories of robots. UniVLA [23] incorporates historical actions into the input prompt, enabling the robot to learn from its own decisions and adapt to dynamic environments. Unlike these methods, our model leverages rich historical frame information to address long-horizon tasks. Information from historical frames is more informative, as it includes not only the actions of robots but also the effects of those actions on the environments, such as occlusion relationships caused by the manipulation of robots.

3 Methodology

Our goal is to develop a VLA model that enables different robots to accurately perform various tasks based on historical information, multi-view images, and language instructions. Specifically, given a long-horizon video input, multi-view images at the current single frame, and a language instruction, the proposed model predicts a temporal action sequence $\{a_t, a_{t+1}, a_{t+2}, ..., a_{t+s}\}$ to drive the robot to complete the corresponding tasks, where s is the number of predicted future

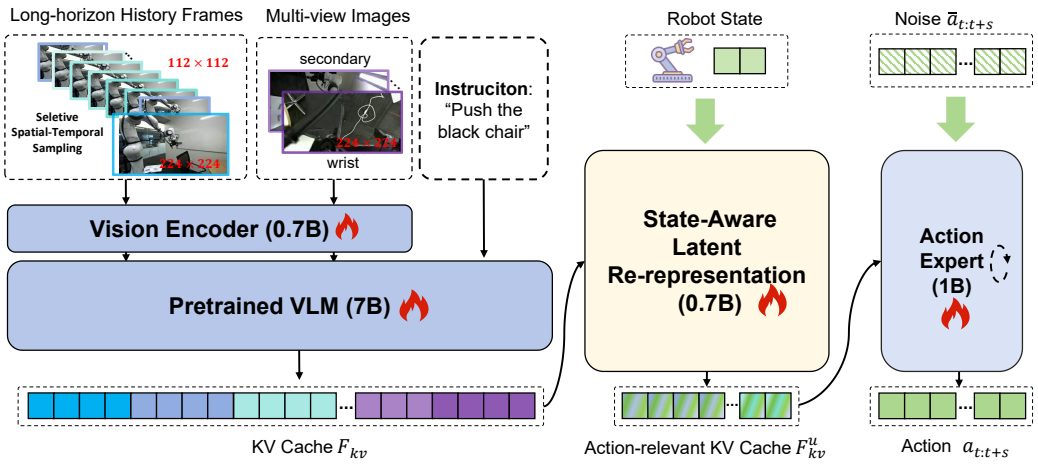


Figure 1: Overview of the proposed model **HALO**. **HALO** has 10 billion trainable parameters and can process long-horizon history frames (max to 8s when FPS is 10). To achieve the modality alignment between the vision-language embeddings and actions, the state-aware latent re-representation fuses the state information from robots and the vision-language embeddings.

steps. As shown in Figure 1, the proposed model consists of three components: the pretrained VLM, the State-Aware Latent Re-representation, and the Action Expert. Pretrained VLM is responsible for selecting the visual tokens most relevant to the language instruction from the long video and multi-view images. The State-Aware Latent Re-representation aligns the vision-language embeddings with actions, and the Action Expert decodes the desired action for the robot from noise based on the aligned embeddings.

3.1 Pretrained VLM for Long-Horizon Video Encoding

Video Encoding. Given a history video sequence $\mathcal{V}$ which contains n frames and is obtained from the primary view:

$$\mathcal{V} = \{V^{\downarrow}_{t-n}, ..., V^{\downarrow}_{t-3}, V^{\downarrow}_{t-2}, V^{\downarrow}_{t-1}, V_t\}, \tag{1}$$

where $V^{\downarrow}$ and V_t denotes the downsampled frames and the t-th frame without downsampling, respectively. The primary view typically refers to a camera mounted at the front of the robot, which faces the task area. It provides the most critical and comprehensive perspective for observing the environment. The vision encoder from the pretrained vision-language model (VLM) is used to extract visual tokens $F_v \in \mathbb{R}^{L_v \times H}$, where L_v is the length of the video tokens and H is the hidden size. Because Qwen2.5-VL [29] is capable of understanding long videos exceeding one hour in duration by integrating dynamic frame rate (FPS) training with absolute time encoding, we choose it as the pretrained VLM. By adapting to varying frame rates, it can better capture the temporal dynamics of video content. To reduce the computational burden, we propose a selective spatial-temporal sampling strategy, which downsamples the resolution of historical frames while preserving high-resolution inputs for the current multi-view observations. Specifically, each frame of the video $\mathcal{V}$ is first resized to 112×112 and then is fed into the vision encoder. Although the resolution of historical frames is lower, their large number allows for complementary information during the feature extraction phase, thereby reducing information loss.

Multi-View Image Encoding. Since the resolution of the image of primary view is relatively low, some information may be lost. Therefore, the multi-view images $\mathcal{V}$ at the current time t:

$$\mathcal{V}_m = \{V_{sec}, ..., V_{wrist}\}, \tag{2}$$

are also fed into the vision encoder to mitigate this information loss, which keep the original size. The images from secondary view V_{sec} refers to images captured from alternative angles (e.g., side or top-down perspectives), which can help supplement the occluded regions from primary view. The wrist view is a camera mounted at the end of the arm of robot, which is near the gripper or tool and offers a close-up, detail, rich perspective that is useful for fine-grained manipulation tasks. Specifically, the

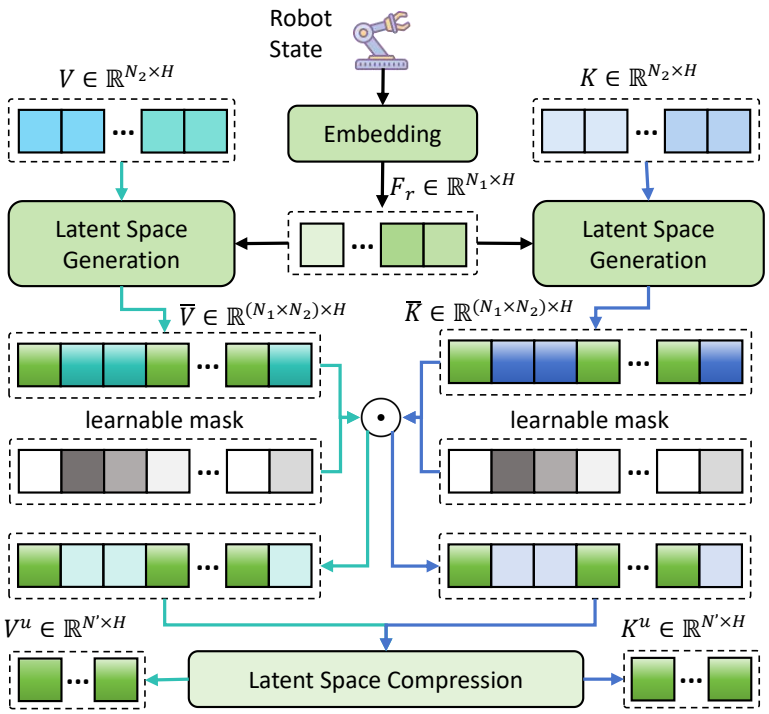


Figure 2: Illustration of the state-aware latent re-representation. The latent space generation calculate the product between each element from state embeddings and each element from the key token K or value token V. The learnable mask determine. The learnable mask determines which information in the latent space is retained and which is suppressed. The latent space compression further compresses the latent space to filter out redundant information.

extracted multi-view tokens $F_m \in \mathbb{R}^{L_m \times H}$ and video tokens F_v are concatenated and then jointly fed into the large language model with the language tokens F_l to perform vision-language joint perception:

$$F_{kv} = \text{VLM}(F_v, F_m, F_l), \tag{3}$$

where F_{kv} is the output KV cache.

3.2 State-Aware Latent Re-representation

In vision-language-action (VLA) model, there exists a significant gap between vision-language embeddings and actions due to the inherently different modalities and representations of high-level semantic information (e.g., language and vision) and low-level motor control signals. This discrepancy makes it challenging for the model to directly translate abstract instructions and visual cues into precise robotic actions. Therefore, effectively aligning these modalities is critical to ensure that the robot can correctly understand the task and perform accurate, goal-directed behaviors. However, current methods [5, 7, 9] suffer from weak alignment between actions and vision-language embeddings. They usually use vision-language embeddings directly as conditions to predict future actions, which may lead to irrelevant information in the embeddings (e.g., background features) misguiding the action generation.

The state information of robots typically includes the joint angles or the position of the end-effector at the current time step, and this modality is naturally aligned with the action space. Therefore, we propose the Re-representation of State-Aware Latent, which leverages the state information of robots to select action-aligned information from the redundant vision-language embeddings. Specifically, the KV cache $F_{kv} \in \mathbb{R}^{L \times 2N_2 \times H}$ output by the pretrained VLM is first split into keys $F_k \in \mathbb{R}^{L \times N_2 \times H}$ and values $F_v \in \mathbb{R}^{L \times N_2 \times H}$, where L, N_2, H denotes the length of tokens, the number of heads and the hidden size, respectively. They are not aligned with the action of the robot, so their representations

need to be updated to achieve alignment. We first propose to generate a larger action-relevant latent space and then search within this space for features similar to actions. Specifically, as shown in Figure. 2, we perform head-wise outer-product fusion between the current state embeddings $F_r \in \mathbb{R}^{N_1 \times H}$ of robots and the vision-language key embeddings, where N_1 denotes the number of heads. Given the key token $K \in \mathbb{R}^{N_2 \times H}$ from F_k, the value token $V \in \mathbb{R}^{N_2 \times H}$ from F_v and state embeddings F_r, the fused representation $\bar{K} \in \mathbb{R}^{(N_1 \times N_2) \times H}$ and $\bar{V} \in \mathbb{R}^{(N_1 \times N_2) \times H}$ are computed, respectively, which capture rich inter-head interactions across modalities. Formally, it is defined as:

$$\begin{aligned} \bar{K}[i, j, :] &= F_r[i, :] \odot K[j, :], \\ \bar{V}[i, j, :] &= F_r[i, :] \odot V[j, :], \end{aligned} \tag{4}$$

where $\odot$ denotes the element-wise product. Then, to extract action-relevant cues from the latent space, we introduce a learnable mask for both the key tokens and value tokens, which adaptively determines how much information to retain. Formally, this process can be written as:

$$\begin{aligned} K' &= M_k \odot K, \\ V' &= M_v \odot V, \end{aligned} \tag{5}$$

where $M_k \in \mathbb{R}^{(N_1 \times N_2) \times H}$ and $M_v \in \mathbb{R}^{(N_1 \times N_2) \times H}$ denotes the learnable mask for key token and value token, respectively. Finally, to further compress the representation space, we propose a latent space compression strategy to obtain re-encoded key embeddings $K^u \in \mathbb{R}^{N' \times H}$ and value embeddings $V^u \in \mathbb{R}^{N' \times H}$, where N' denotes the new number of heads.

3.3 Action Expert for Action Prediction

We use a conditional flow matching action expert [5] for fine-grained end-effector action generation, which consists of a series of Transformer self-attention layers from pretrained large language model. It takes the aligned vision-language embeddings as input condition to generate future multi-step actions $\{a_t, a_{t+1}, ..., a_{t+s}\}$ and predicts actions through the progressive fusion of these embeddings with noise. During inference, the Action Expert performs multiple denoising steps to progressively decode the actions from noise.

4 Experiments

4.1 Implementation Details

Our HALO model is pretrained with OXE [30] and AgiBot dataset [11], consisting of 1.1 million real-world robot episodes on a cluster of 32 A100 40G GPUs for 14 days. The VLM part of our model is initialized from Qwen2.5VL-7B [29], and the full 10B model is trained in an end-to-end fashion. Specifically, we use FSDP as our distributed training framework with hybrid sharding strategy. Gradient checkpointing is used to reduce memory usage per batch. A gradient accumulation step of 4 is utilized to boost batch size to 1280. We use LeRobot Dataset as our unified dataset format. We further conduct inference experiment. The result shows that our model can support up to 30Hz of control frequency on NVIDIA Geforce RTX A6000 GPUs.

4.2 Main Results

To demonstrate the effectiveness of our proposed model, we evaluate the model HALO across multiple widely-used simulation benchmarks (including SIMPLER [31], LIBERO [32] and CALVIN [33]) and real-world scenarios. Besides, we further categorize the tasks into **Single-Step** and **Multi-Step** tasks. The former requires executing only one atomic action (e.g., "pick", "put"), while the latter involves a sequence of actions (e.g., "open" followed by "place").

4.2.1 Manipulation Benchmark on SIMPLER

The SIMPLER [31] evaluation environment aims to bridge the real-to-sim control and visual gap. It replicates real-world scenarios on the Google Robot and WidowX Robot. There are two real-to-sim evaluation setups: **Visual Matching**, which aims to reduce the visual appearance gap between real environments and raw simulation by overlaying real-world images onto simulation backgrounds,

Table 1: Comparison of our approach with existing VLA models across four tasks in two SIMPLER settings on the Google robot.

Google Robot	Method	Single-Step	Multi-Step			Average
		Pick Coke Can	Move Near	Open/Close Drawer	Open Top Drawer and Place Apple	
Visual Matching	RT-1 [34]	85.7%	44.2%	73.0%	6.5%	52.4%
	RT-1-X [10]	56.7%	31.7%	59.7%	21.3%	42.4%
	RT-2-X [10]	78.7%	77.9%	25.0%	3.7%	46.3%
	Octo-Base [20]	17.0%	4.2%	22.7%	0.0%	11.0%
	OpenVLA [4]	18.0%	56.3%	63.0%	0.0%	34.3%
	π_0 [5]	87.3%	35.0%	72.6%	16.0%	52.7%
	HALO (Ours)	-	**44.6%**	-	**19.4%**	-
Variant Aggregation	RT-1 [34]	89.8%	50.0%	32.3%	2.6%	43.7%
	RT-1-X [10]	49.0%	32.3%	29.4%	10.1%	30.2%
	RT-2-X [10]	82.3%	79.2%	35.3%	20.6%	54.4%
	Octo-Base [20]	0.6%	3.1%	1.1%	0.0%	1.2%
	OpenVLA [4]	60.8%	67.7%	28.3%	1.2%	39.3%
	π_0 [5]	85.2%	40.8%	42.1%	15.9%	46.0%
	HALO (Ours)	-	**40.3%**	-	-	-

Table 2: Comparison of our approach with existing VLA models across four tasks in the SIMPLER (Visual Matching) setting on the WidowX robot.

WidowX Robot	Method	Multi-Step				Average
		Put Spoon on Towel	Put Carrot on Plate	Stack Green Block on Yellow Block	Put Eggplant in Yellow Basket	
Visual Matching	RT-1-X [34]	0.0%	4.2%	0.0%	0.0%	1.1%
	Octo-Base [20]	15.8%	12.5%	0.0%	41.7%	17.5%
	Octo-Small [20]	41.7%	8.2%	0.0%	56.7%	26.7%
	OpenVLA [4]	4.2%	0.0%	0.0%	12.5%	4.2%
	π_0 [5]	62.5%	66.7%	25.0%	12.5%	41.7%
	SpatialVLA [35]	16.7%	25.0%	29.2%	100%	42.7%
	CogACT [7]	71.7%	50.8%	15.0%	67.5%	51.3%
	HALO (Ours)	54.2%	41.7%	54.2%	79.2%	57.3%

and **Variant Aggregation**, which creates different simulation environment variants (e.g., different backgrounds, lightings, distractors, table textures) based on Visual Matching. We compare our model with the latest state-of-the-art VLA models under two evaluation settings. Table 1 summarizes the results of different VLA methods on two evaluation settings of the Google robot dataset. Our model achieves state-of-the-art performance in both settings, with 66.4% on **Visual Matching** and 65.6% on **Variant Aggregation**. Specifically, compared to π_0, our model achieves substantial improvements on multi-step tasks, outperforming it by 16% and 15% on **Visual Matching** and **Variant Aggregation**, respectively. Moreover, despite having fewer parameters (10B vs 55B), our model significantly surpasses the closed-source RT-2-X in terms of success rate.

Table 2 summarizes the results of different methods on the WidowX robot. Our model also achieves the highest success rate, significantly outperforming other approaches. The tasks for this robot often involve multiple atomic actions and can thus be considered as multi-step tasks. For example, "put spoon on towel" requires first executing a pick action, followed by a put action. As shown in Table 2, our method achieves an overall improvement of 15.6% over π_0, demonstrating its ability to effectively extract task-relevant motion cues from historical information for more accurate action generation. Moreover, we observe that our model is capable of self-correction by leveraging historical context. For instance, when performing the "stack block" task, if it fails to grasp the green block on the first attempt, it continues to retry, with each subsequent attempt becoming more accurate.

Table 3: Comparison of our approach with existing VLA models on the LIBERO simulation environments.

Method	Single-Step	Multi-Step			
	LIBERO-Goal	**LIBERO-Object**	**LIBERO-Spatial**	**LIBERO-Long**	**Average**
Diffusion Policy [18]	68.3%	92.5%	78.3%	50.5%	72.4%
Octo [20]	84.6%	85.7%	78.9%	51.1%	75.1%
OpenVLA [4]	79.2%	88.4%	84.7%	53.7%	76.5%
TraceVLA [6]	75.1%	85.2%	84.6%	54.1%	74.8%
RDT [21]	68.2%	77.8%	60.2%	29.0%	58.8%
π_0 [5]	94.0%	97.8%	91.4%	85.4%	92.2%
HALO (Ours)	94.3%	97.4%	92.0%	85.6%	92.3%

4.2.2 Manipulation Benchmark on LIBERO

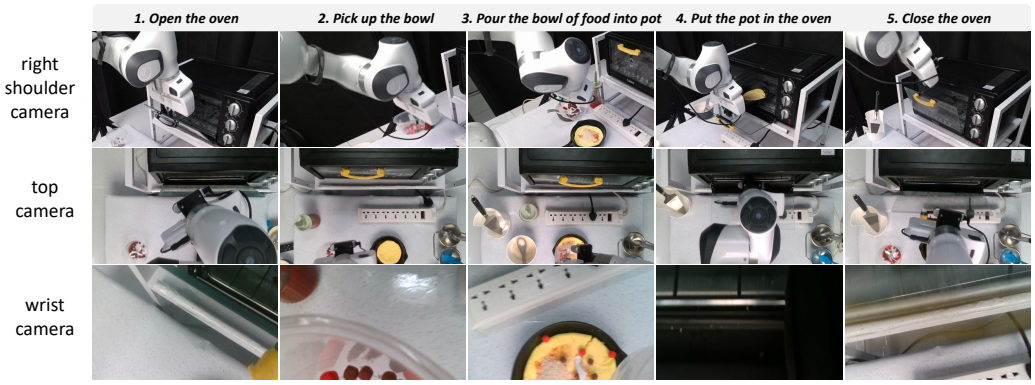
The LIBERO [32] benchmark consists of four task suites, which are designed to study lifelong learning in robotic manipulation. We perform experiments on four task suites, each comprising 10 tasks with 50 human-teleoperated demonstrations. Specifically, **LIBERO-Spatial**, **LIBERO-Object** and **LIBERO-Goal** evaluate the understanding of the spatial relationships, object types and different task-oriented behaviors, respectively. **LIBERO-Long** test the ability to generalize the long-horizon tasks with different objects, layouts and goals. Our model is fine-tuned on the mixed LIBERO dataset for $30k$ steps with a batch size of 128. Additionally, to ensure a fair comparison, we reproduce the results of π_0 on the LIBERO benchmark. Since LIBERO-Object, LIBERO-Spatial, and LIBERO-Long contain multi-step instructions, we categorize them as **Multi-Step** dataset. Table 3 compares the performance of different VLA models on the LIBERO dataset. Our model achieves the highest average success rate, surpassing existing state-of-the-art methods. Specifically, on the Multi-Step datasets, our model outperforms π_0 by 0.2% on LIBERO-Long and 0.6% on LIBERO-Spatial, demonstrating that historical information can effectively guide the robot to perform accurate actions.

Table 4: Comparison of our approach with existing VLA models on the CALVIN benchmark. We report the success rates as well as the average number of completed tasks per evaluation sequence (with a maximum of 5 tasks).

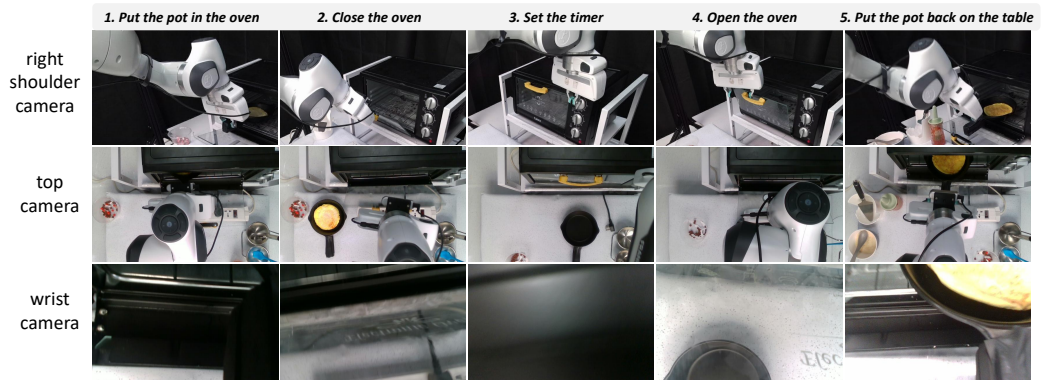
Method	**Experiment**	**Tasks completed in a row**					**Avg. Len.**
		1	2	3	4	5	
MCIL [36]	ABC→D	0.304	0.013	0.002	0.000	0.000	0.31
Diffusion Policy [18]	ABC→D	0.402	0.123	0.026	0.008	0.000	0.56
RT-1 [34]	ABC→D	0.533	0.222	0.094	0.038	0.013	0.90
HULC [37]	ABC→D	0.418	0.165	0.057	0.019	0.011	0.67
MT-R3M [38]	ABC→D	0.529	0.234	0.105	0.043	0.018	0.93
RoboFlamingo [39]	ABC→D	0.824	0.619	0.466	0.331	0.235	2.47
GR-1 [40]	ABC→D	0.854	0.712	0.596	0.497	0.401	3.06
3D Diffuser Actor [41]	ABC→D	0.938	0.803	0.662	0.533	0.412	3.35
UniVLA [23]	ABC→D	0.955	0.858	0.754	0.669	0.565	3.80
π_0 [5]	ABC→D	0.842	0.614	0.442	0.316	0.216	2.43
HALO (Ours)	ABC→D	-	-	-	-	-	-

4.2.3 Manipulation Benchmark on CALVIN

CALVIN [33] is a challenging simulated benchmark and aims to learn language-conditioned policy for long-horizon robot manipulation tasks. It contains 34 tasks and the environment use the Franka Emika Panda robot with a parallel-jaw gripper to perform the task. We conduct experiments on the subset **ABC→D**, where A, B, C, and D represent different environments with variations in desk colors and object configurations. In this setting, **ABC→D** denotes training on data from environments A, B, and C, and testing on environment D, which serves as a zero-shot evaluation. The evaluation consists of a set of 1,000 unique instruction chains, each comprising five consecutive tasks, designed to comprehensively assess the generalization capability of the policy. Our model is finetuned on



(a) simple scenario: the target (pot) is visible in the view



(b) the target (pot) is occluded inside the oven and requires long-term memory for accurate localization

Figure 3: Real-world evaluation of the long-horizon task *"Heat the Food"* using the Franka robot. The task includes two scenarios: (a) the target object (a pot) remains visible throughout the process, and (b) the target becomes occluded in the final step after being placed in the oven at the beginning.

the CALVIN training set for $300k$ steps with a batch size of 32. Meanwhile, we also finetune π_0 on CALVIN using the same settings for a fair comparison. The results in Table 4 demonstrate the state-of-the-art performance of our method on long-horizon tasks. Our approach achieves a 76.1% success rate in completing all five tasks, surpassing the previous best methods, RoboVLM by 5.7% and 3D Diffusion Actor by 34.9%. Besides, under the same training settings, our model significantly outperforms π_0, exceeding it by 2.2 in terms of the average number of completed tasks. Notably, by effectively leveraging historical information, our model is capable of handling complex, long-horizon manipulation tasks.

Table 5: Comparison of our approach with existing VLA models in real-world scenarios with the Franka Robot.

	Single-Step			**Multi-Step**			
Method	T1	T2	T8	T2 → T3 → T4	T5 → T2 → T4 → T7 → T8	T5 → T7 → T8 → T6 → T5 → T9	**Average**
OpenVLA [4]	0	27.3	0	0	0	0	5.8
π_0 [5]	70.5	72.5	64.3	20.2	25.7		50.6
Ours	-	-	-	-	-		-

4.2.4 Real-World Evaluation with Franka Robot

Self-collected Data. We conduct experiments on the Franka robot, which has 7 DoFs and is equipped with a 1-DoF gripper.

Training and Evaluation Details. The implementation details trained on the real-world dataset are consistent with the fine-tuning in the simulation environment. Besides, we define 8 primitive tasks for **Single-Step** and **Multi-Step** for real-world evaluation: *Pick Brush* (**T1**), *Pick Bowl* (**T2**), *Move Bowl* (**T3**), *Pour Food* (**T4**), *Open Owen* (**T5**), *Set the Timer* (**T6**), *Put the Pot in the Oven* (**T7**), *Close Owen* (**T8**) and *Put the Pot on the Table* (**T9**). To evaluate the ability on long-horizon scenes, we design three multi-step tasks: *Pour Food into Pot*: $T2 \rightarrow T3 \rightarrow T4$, and *Heat the food*: $T5 \rightarrow T2 \rightarrow T4 \rightarrow T7 \rightarrow T8$, and $T5 \rightarrow T7 \rightarrow T8 \rightarrow T6 \rightarrow T5 \rightarrow T9$.

Results. Table 5 summarizes the results of our model compared with OpenVLA and π_0. Our results demonstrate state-of-the-art performance on both single-step and multi-step tasks. Specifically, our model achieves a 21% higher success rate than π_0 on the five-step task *"Heat the Food"*, demonstrating a clear advantage in handling long-horizon tasks.

Table 6: Impact of each component. **Pretrain**, **MF** and **SALR** denotes the weights pretrained on robot dataset, the multiple frames and the State-Aware Latent Re-representation, respectively.

Frozen VLM	MF	SALR	Put Spoon on Towel	Put Carrot on Plate	Stack Green Block on Yellow Block	Put Eggplant in Yellow Basket	Average
	✓						
		✓					
	✓	✓					
✓	✓	✓					

4.3 Ablation Study

We conduct ablation experiments on the WidowX robot from the SIMPLER simulated environment and report the average manipulation accuracy. **MF** means the multiple historical frames is use. **Pretrain** means the HALO is firstly pretrained on large-scale robot dataset and finetuned on the WidowX robot. **SALR** means the state-aware latent re-representation. Note that without using **MF**, only the current frame is used. Without **SALR**, a simple MLP is applied to convert the number of heads. As shown in Table 6, the overall manipulation success rate improves significantly when **MF** is used. When both $\mathbf{LA}^3$ and **MF** are applied, the model achieves the best performance. These results highlight the importance of leveraging historical frame information and aligning vision-language embeddings with actions.

5 Conclusion

In this paper, we propose **HALO**, a vision-language-action (VLA) model designed to address the challenges of long-horizon robotic manipulation. We propose to use the Qwen2.5-VL to effectively process the historical frames and capture the long-dependencies. To balance the complexity and performance, we further design a selective spatial-temporal sampling strategy to fuse the long historical frames and current multi-view images. Besides, to bridge the modality gap between the actions and vision-language embeddings, we propose the state-aware latent re-representation to fuse their features and then use the aligned embeddings to guide the prediction of future actions. Extensive experiments demonstrate that our model outperforms existing VLA models in task performance, with greater advantages in long-horizon tasks.

References

[1] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[3] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[4] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control, 2024. *URL https://arxiv. org/abs/2410.24164*.

[6] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

[7] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[8] Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, et al. Diffusion-vla: Scaling robot foundation models via unified diffusion and autoregression. *arXiv preprint arXiv:2412.03293*, 2024.

[9] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

[10] Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

[11] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[12] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[13] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

[14] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

[17] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[18] 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. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[19] Moritz Reuss, Jyothish Pari, Pulkit Agrawal, and Rudolf Lioutikov. Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning. *arXiv preprint arXiv:2412.12953*, 2024.

[20] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[21] 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. *arXiv preprint arXiv:2410.07864*, 2024.

[22] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024.

[23] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[24] Stephen James and Andrew J Davison. Q-attention: Enabling efficient learning for vision-based robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):1612–1619, 2022.

[25] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[26] Yunhai Feng, Jiaming Han, Zhuoran Yang, Xiangyu Yue, Sergey Levine, and Jianlan Luo. Reflective planning: Vision-language models for multi-stage long-horizon robotic manipulation. *arXiv preprint arXiv:2502.16707*, 2025.

[27] Shichao Fan, Quantao Yang, Yajie Liu, Kun Wu, Zhengping Che, Qingjie Liu, and Min Wan. Diffusion trajectory-guided policy for long-horizon robot manipulation. *arXiv preprint arXiv:2502.10040*, 2025.

[28] Jessica Borja-Diaz, Oier Mees, Gabriel Kalweit, Lukas Hermann, Joschka Boedecker, and Wolfram Burgard. Affordance learning from play for sample-efficient policy learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6372–6378. IEEE, 2022.

[29] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[30] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[31] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

[32] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[33] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[34] 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.

[35] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

[36] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

[37] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters*, 7(4):11205–11212, 2022.

[38] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[39] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.

[40] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.

[41] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

A Sec.1