Viewpoint Matters: Dynamically Optimizing Viewpoints with Masked Autoencoder for Visual Manipulation

Pengfei Yi[1], Yifan Han[1], Junyan Li[1], Litao Liu[2], and Wenzhao Lian[1]

***Abstract*— Robotic manipulation continues to be a complex challenge, with imitation learning (IL) offering an effective way for robots to learn tasks from expert demonstrations. Current IL methods typically rely on fixed camera setups—either multi-camera systems, which may introduce redundant or noisy data, or single-camera systems, which suffer from limited viewpoints, constraining task performance. Inspired by human active perception, where humans dynamically adjust their viewpoint to capture the most relevant and least noisy information, we propose MAE-Select, a novel framework for active viewpoint selection in single-camera robotic systems. MAE-Select fully leverages pre-trained multi-view masked autoencoder representations and dynamically selects the next most informative viewpoint at each time chunk without requiring labeled viewpoints. This plug-and-play approach enhances learning efficiency and task performance. Extensive experiments demonstrate that MAE-Select improves the capabilities of single-camera systems and, in some cases, even surpasses multi-camera setups. Project will be available at https://sites.google.com/view/mae-select.**

I. INTRODUCTION

Robotic manipulation is a core challenge in robotics, critical to applications ranging from industrial automation to healthcare. Imitation Learning (IL) [1], [2], [3], [4], [5] has become a leading approach for enabling robots to learn complex tasks through expert demonstrations. Recently, advances in deep generative models [6], [7], such as variational autoencoders (VAEs) [8] and diffusion models [9], have empowered IL by allowing robots to process high-dimensional sensory inputs, such as images, leading to promising results in robotic manipulation [10], [11], [12], [13].

However, most current IL methods rely on fixed camera setups, either single or multiple cameras, which pose significant limitations. In fixed single-camera setups [14], [15], though practical and cost-effective, robots face challenges due to the limited field of view, which may obstruct critical parts of the environment or objects, negatively impacting task performance. Multi-camera setups, while designed to provide more comprehensive scene coverage, introduce their own complexities: the abundance of redundant or irrelevant information can overwhelm learning algorithms and decrease efficiency. As shown in Sec.IV, these passive static multi-view setups do not always provide the most task-relevant information, leading to suboptimal decision-making.

In contrast, humans dynamically adjust their viewpoints while performing tasks. By actively moving our head and neck, we naturally seek the most informative, least noisy perspectives that are most relevant to the task at hand. Inspired by this human capability, we propose shifting from passive, static perception to **active perception**, where the viewpoint is dynamically adjusted throughout the task to optimize information intake. In a practical robotic setting, this could be embodied by a humanoid robot moving its neck and head to capture the most task-relevant views in real-time. In this paper, we focus on the feasibility of active viewpoint selection for robotic manipulation as an initial exploration along this direction.

To this end, we introduce **MAE-Select**, a framework designed to actively select optimal viewpoints for single-camera robotic setups. MAE-Select first fully utilizes the powerful pre-trained representations from the multi-view masked autoencoders (MAEs) [16], [17], leveraging its complete encoder-decoder architecture to obtain multi-view representations. Unlike prior works that focus on fixed viewpoints [18], MAE-Select dynamically predicts the next better viewpoint based on the current chunk of visual and action information. Crucially, this viewpoint selection is learned solely through imitation learning, requiring no manual labels for optimal views. Moreover, this innovative mechanism is a plug-and-play solution, making it easy to integrate into various existing setups, and demonstrates significant potential in advancing single-camera robotic manipulation.

Our key contributions are as follows:

- We propose MAE-Select, a novel plug-and-play viewpoint selection mechanism that dynamically selects the next optimal viewpoint at each time chunk.
- We present an imitation learning framework that fully utilizes pre-trained representations from a multi-view masked autoencoder for manipulation.
- We demonstrate through experiments across various scenarios and tasks that MAE-Select significantly enhances manipulation efficiency and accuracy in single-camera setups, even outperforming multi-camera systems in certain tasks.

II. RELATED WORK

A. Imitation learning for Manipulation

Imitation learning, which enables agents to learn tasks by observing and mimicking expert actions, has been widely applied to manipulation tasks [19], [20], [21], [22], [2], [23], [24]. Behavioral cloning [1] is a foundational approach that treats imitation as a supervised learning problem, mapping observations directly to actions. Recently, significant

*Corresponding author: Wenzhao Lian

[1] Pengfei Yi, Yifan Han, Junyan Li, and Wenzhao Lian are with Institute of Automation, Chinese Academy of Sciences. {yipengfei2024, hanyifan2024, lijunyan2024, wenzhao.lian}@ia.ac.cn

[2] Litao Liu is with CoreNetic.ai. liulitao6688@gmail.com

Manuscript 2428 submitted to 2025 IEEE International Conference on Robotics and Automation (ICRA). Received September 15, 2024.

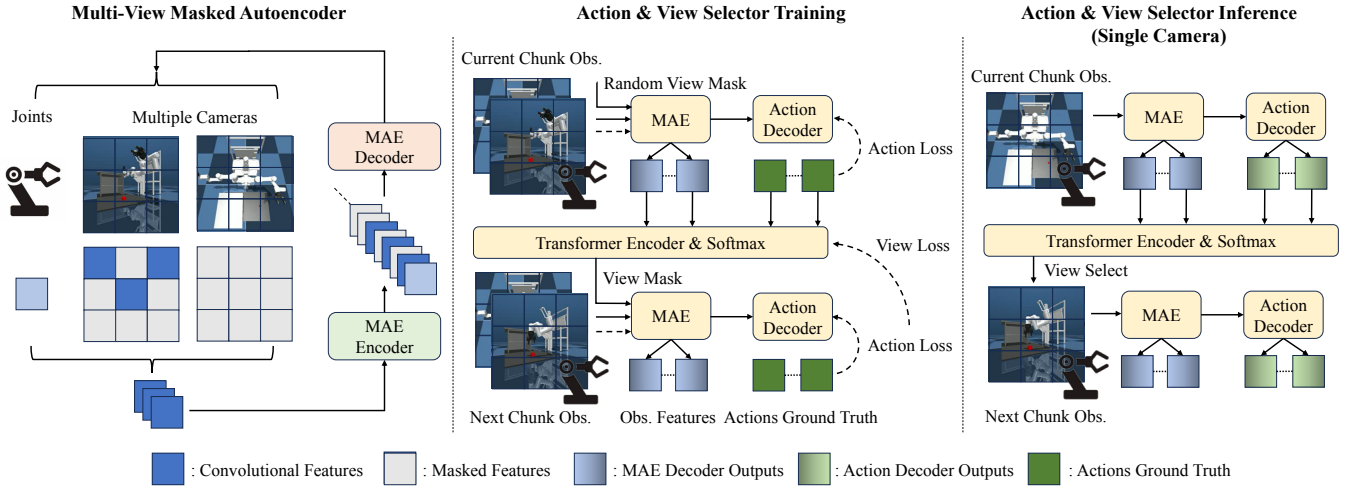

Fig. 1. Illustration of our proposed method. **Left** depicts the pre-training stage of the multi-view masked autoencoder with joint embeddings. **Middle** illustrates the training process of our framework using imitation learning. **Right** demonstrates how the framework operates during inference.

advancements have been made in imitation learning (IL) for robotic manipulation. With the advent of deep learning, there has been a surge in methods leveraging computer vision techniques to extract information from images [25], [26], [27], [11], [10], virtual reality (VR) [28], and 3D point clouds [29], [30]. For instance, ACT [11] introduced action chunking with transformers, enhancing both accuracy and efficiency. Diffusion Policy [10] represents a robot's visuomotor policy as a conditional denoising diffusion process, enabling stable learning of complex manipulation tasks from expert demonstrations. However, most of these methods rely on fixed camera setups. In this work, we focus on the critical role of viewpoints in manipulation and aim to unlock the full potential of single-camera applications to achieve superior results.

B. Unsupervised representation learning

Unsupervised representation learning [31], [32], [33] aims to learn useful representations or features from data without using any labeled examples. A leading method in this area is the Masked Autoencoder (MAE) [16], which operates by randomly masking portions of the input and training the model to reconstruct the missing parts. This technique has shown outstanding performance, particularly in computer vision tasks [34], [35], [17]. In the field of robotics, several approaches [36], [37], [18] have applied MAE to enhance image representation, leading to improvements in manipulation. Among these, MV-MWM [18] utilizes MAE with multi-view data to enhance visual robotic manipulation, achieving notable performance in single-view control scenarios. However, its focus is primarily on enhancing the performance of a fixed single viewpoint using auxiliary data from other viewpoints, without exploring alternative viewpoints.

III. METHOD

We present MAE-Select, a plug-and-play framework that selects the next better viewpoint for visual robotic manipulation. It learns multi-view representations through masked autoencoder and fully utilizes the whole encoder-decoder structure of the autoencoder for the action decoder and viewpoint selector in contrast to the previous methods [36], [37], [18]. We first introduce how to learn multi-view representations in Sec.III-A. Then, we describe how to fully utilize the pre-trained masked autoencoder for manipulation in Sec.III-B. Finally, we present the way to learn better view point choices in conjunction with imitation learning in Sec.III-C. The overview of our method is shown in Fig.1.

Let $O_t = \{o_t^{v_1}, ..., o_t^{v_i}\}$ be an image set from multiple viewpoints, s_t be the joint positions of the robot and $a_{t,T} = \{a_t, ..., a_{t+T-1}\}$ be the action chunking with the window size of T at timestep t, where v_i represents a viewpoint respectively and the length of O_t represents the number of available viewpoints.

A. Multi-View Representation

Our aim is to learn multi-view representations and reconstruct other viewpoints from the one that is important to select the next better viewpoint. The core idea is to pretrain a multi-view masked autoencoder with view-masking and joint embeddings to recover the randomly masked viewpoints and image patches.

Random Masking. The vanilla masked autoencoder [16] masks random image pixel patches. As pixel patch masking makes it difficult for the model to learn fine-grained details, we randomly mask the convolutional features in a similar manner to [37], [18]. Specifically, given an image $o_t^{v_i} \in \mathbb{R}^{224\times224\times3}$, it is first fed into 4 convolutional layers. The output from the final convolutional layer produces feature map $\bar{o}_t^{v_i} \in \mathbb{R}^{14\times14\times d_{in}}$ and it is flattened into $\bar{\bar{o}}_t^{v_i} \in \mathbb{R}^{196\times d_{in}}$, which then undergoes the masking process. The observations from each viewpoint are separately processed with shared parameters. This approach allows the model to focus on learning more nuanced and detailed features, as the convolutional features retain more spatial information compared to

Manuscript 2428 submitted to 2025 IEEE International Conference on Robotics and Automation (ICRA). Received September 15, 2024.

raw pixel patches. Given a mask ratio of m, a proportion m of the feature map $\bar{\bar{o}}_t^{v_i}$ is randomly selected for masking. We also adopt view-masking [18] that randomly masks whole viewpoints to make the masked autoencoder to learn the cross relationship among these viewpoints.

Multi-View Masked Autoencoder. For each viewpoint, we add fixed 2D sin-cos position embeddings to the features to encode spatial information. Additionally, we incorporate learnable 1D parameters representing each viewpoint to the features. Initially, the unmasked features from all viewpoints are processed through a series of vision transformer (ViT) [38] layers to generate mixed features. To assist in the reconstruction, joint state information with learnable position embeddings is incorporated. Specifically, a set of mask tokens is concatenated with the encoded features and joint features. The decoder then processes these through ViT layers and linearly projects them into pixel patch predictions and state predictions. The process can be presented as,

$$\begin{aligned} \text{Masking: } & h_t^m \sim p^{mask}(h_t^m \mid \{\bar{\bar{o}}_t^{v_i}\}_{v_i \in V}, m) \\ \text{ViT Encoder: } & z_t^m = f_\phi(h_t^m) \\ \text{ViT Decoder: } & \{\hat{o}_t^{v_i}\}_{v_i \in V}, \hat{s}_t = g_\phi(z_t^m, s_t) \end{aligned} \tag{1}$$

We train the model to reconstruct both pixels and states, optimizing the model parameters ϕ by minimizing the prediction error,

$$L^{mae}(\phi) = MSE(\hat{o}_t^m, o_t^m) + L_1(\hat{s}_t, s_t) \tag{2}$$

B. Imitation Learning

Unlike previous methods [36], [37], [18], we utilize the whole encoder-decoder of the masked autoencoder, excluding the linear head at the end, to produce multi-view features and it can be integrated into various types of action decoders. By default, we employ the transformer-based diffusion policy [10] as our decoder due to its robustness in handling complex tasks. With a dataset $\mathscr{D} = \{O_t, s_t, a_t\}_{t=0\cdots T_m}^N$, where T_m represents the maximum length of time steps and N represents the total number of trajectories, our objective is to learn a policy $\pi_\theta(a_{t,T}|s_t, O_t)$ through imitation learning, ensuring that the generated actions $\hat{a}_{t,T}$ closely match those provided in the expert demonstrations. Specifically, given the observations O_t and s_t, we have a 50% probability of not masking O_t. Otherwise, one viewpoint is randomly retained, while the others are masked, as $O_t^m = o_t^v$. Then they are fed into the pre-trained masked autoencoder,

$$C_t = g_\phi(f_\phi(O_t^m), s_t) \tag{3}$$

We also use a Denoising Diffusion Probabilistic Model (DDPM) to approximate the conditional distribution $p(a_{t,T} \mid C_t)$. During training, following [10], we randomly select a denoising iteration k from the uniform distribution $\mathbb{U}(1,K)$. For this iteration, we sample a random noise ε^k with an appropriate variance. The policy predicts the noise ε^k and is updated using the loss function,

$$L^{action}(\theta) = SmoothL1_{\beta=1.0}(\varepsilon^k, \pi_\theta(C_t, a_{t,T}^0 + \varepsilon^k, k)) \tag{4}$$

To ensure that the pre-trained masked autoencoder maintains the predictive power of the image while enhancing feature extraction, we update the whole model by combining the reconstruction loss with the action loss,

$$L^{IL} = L^{action}(\theta) + \alpha L^{mae}(\phi) \tag{5}$$

C. Next Better Viewpoint Selection

Although we have the ground truth of actions for training the action decoder, obtaining the ground truth for the optimal viewpoint is challenging. This difficulty arises from the fact that the optimal viewpoint depends on the task and context, making it impractical to define a "correct" viewpoint. Given this, we take an approach inspired by recurrent neural networks (RNNs) [39], [40] to iteratively refine viewpoint selection during imitation learning as shown in Fig.1. We randomly initialize the viewpoint for the first time chunk, and the selector model outputs the better viewpoint for the second time chunk. The objective is to minimize the action prediction error of the second time chunk. Using gradient descent, we update the viewpoint selection network to refine its ability to choose optimal viewpoints implicitly.

Specifically, let $\mathscr{D}_t$ and $\mathscr{D}_{t+T}$ represent two consecutive time chunks, with O_t being the observations at the current time chunk and O_{t+T} the observations at a future time chunk. We start by assuming an initial random viewpoint, denoted as $O_t^m = o_t^v$, where v refers to one of the candidate viewpoints. These observations are then passed through the multi-view representation models as described in Sec.III-A, extracting multi-view observation features. Together with the actions ground truth $a_{t,T}$, these features are input into the viewpoint selection model, which consists of two layers of transformer encoder $TF(\cdot)$ with a [CLS] token and SoftMax activation. The model then outputs the next, better viewpoint probability p_{t+T}^v for the subsequent time chunk,

$$p_{t+T}^v = SoftMax(TF(C_t, a_{t,T})) \tag{6}$$

Once the probability distribution p_{t+T}^v is obtained, the next viewpoint $\hat{v}_{t+T}$ for the subsequent time chunk $\mathscr{D}_{t+T}$ is chosen based on the highest probability, i.e.,

$$\hat{v}_{t+T} = arg\max_v p_{t+T}^v \tag{7}$$

The observation for the next time chunk is updated as $O_{t+T}^m = o_{t+T}^{\hat{v}_{t+T}}$. According to Sec.III-B, the action loss for the next time chunk L_{t+T}^{action}, is computed. The loss for the viewpoint selection model is then defined as:

$$L^{view} = \sum_v (p_{t+T}^v \cdot L_{t+T}^{action}) \tag{8}$$

Since p_{t+T}^v is obtained using a SoftMax activation, L^{view} is equal to L_{t+T}^{action}. Thus, the viewpoint selection loss L^{view} directly ties the quality of the selected viewpoint to the corresponding action loss, encouraging the model to optimize viewpoints that minimize future action errors.

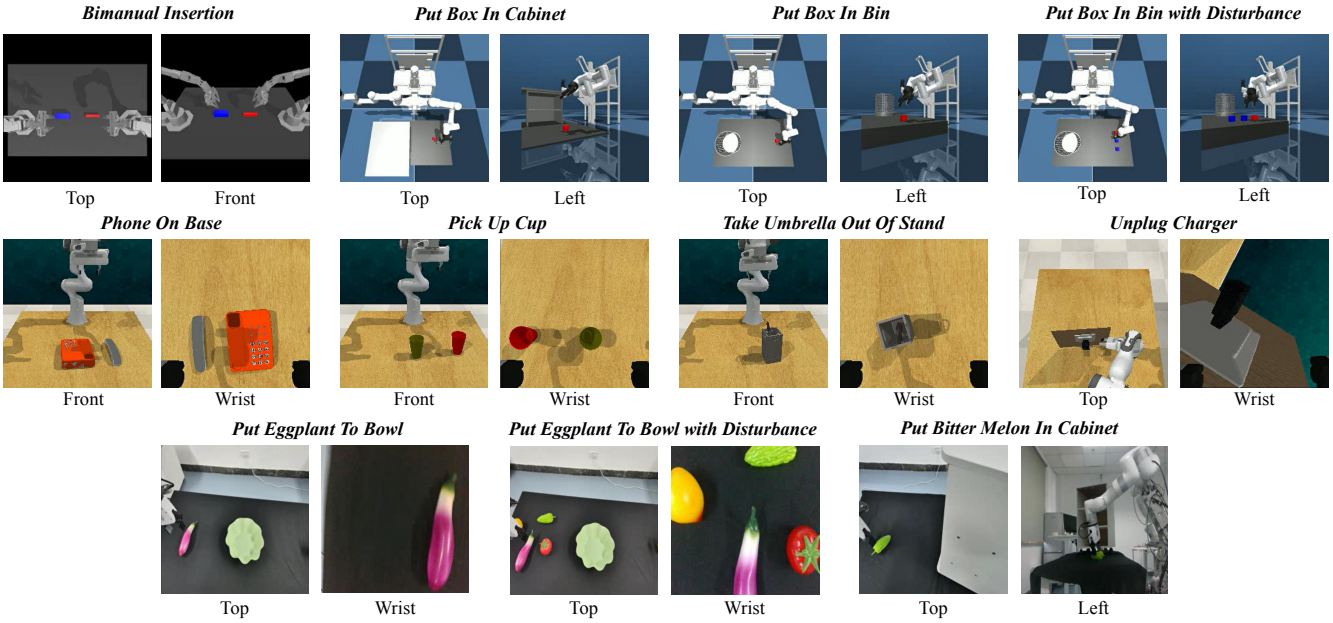

Fig. 2. The viewpoint settings for various robotic tasks, showcasing the viewpoints used to evaluate performance across different simulation and real-world scenarios.

During inference, we randomly set the initial viewpoint and use the initial state as the action with initial observations to predict the optimal viewpoint for the first time chunk. For subsequent chunks, the actions generated by the action decoder are fed into the viewpoint selection module to determine the optimal viewpoint for the next time chunk.

IV. EXPERIMENTS

We evaluate MAE-Select across 3 challenging scenarios and 8 demanding tasks in simulation, including simulations in ACT [11], RLBench [41], and our designed robot in MuJoCo [42], along with three real-world tasks.

A. Tasks

The 11 tasks cover various scenarios, providing a comprehensive evaluation of our method. The viewpoint settings are shown in Fig.2. For simulations in ACT [11], we focus on the *Bimanual Insertion* task, where the robot's arms pick up a socket and a peg for a mid-air insertion. In RLBench [41], we select four varied tasks as follows. *Phone On Base*: The robot needs to pick up and place a phone onto its base. *Pick Up Cup*: The robot needs to grasp and lift a specific cup within several cups. *Unplug Charger*: the robot is tasked with removing a charger from a socket, involving careful manipulation to avoid damaging delicate components. *Take Umbrella Out Of Stand*: This task requires the robot to pick up a small umbrella. We also design three customized tasks for our robot in MuJoCo [42]. *Put Box In Cabinet*: The robot needs to pick up a box and place it inside a cabinet, requiring precise spatial reasoning. *Put Box In Bin*: In this task, the robot places a box into a bin, testing its ability to interact with constrained spaces. *Put Box In Bin with Disturbance*: Similar to the previous task, but the robot must select a specific block from multiple blocks. For real-world evaluation, we designed three tasks similar to the simulation setups but involving real objects: *Put Bitter Melon In Cabinet*, *Put Eggplant To Bowl*, and *Put Eggplant To Bowl with Disturbance*. Our robot consists of a 7DOF Ufactory xarm 7 robotic arm and a parallel-jaw gripper; our camera setup includes two statically mounted (top and left) and one wrist-mounted Realsense D435 cameras.

B. Implementation

We base our implementation on the architecture of the diffusion policy [10]. The action space corresponds to the joint angles of the robot arm, while the image observations have a resolution of 224×224 with a patch size of 16. Our masked autoencoder utilizes a 12-layer ViT [38] for encoder and an 8-layer ViT for decoder, with an embedding dimension of 512. During pretraining, we use a batch size of 128 over 100 epochs. For RLBench and real-world tasks, the time chunk size is set to 20, with a total of 600 epochs. For other tasks, the time chunk size is 100, with 1,000 epochs in total. In the case of MAE-Select, the batch size is set to 32, and the training undergoes two stages, where only the imitation component is trained during the first half of the epochs, and both imitation and view selection components are trained in the second half. For Diffusion Policy, the batch size is set to 64, and all other parameters follow those outlined in the original work [10]. For each method, we evaluate the best-performing checkpoints from the last three evaluated at 100-intervals, with 50 environment initializations in simulation. All models were trained and tested on NVIDIA RTX 4090 GPUs.

Manuscript 2428 submitted to 2025 IEEE International Conference on Robotics and Automation (ICRA). Received September 15, 2024.

TABLE I

RESULTS OF COMPARISON EXPERIMENT. * REPRESENTS WITH DISTURBANCE. BOLD AND UNDERLINED FONTS MEAN THE BEST AND SECOND-BEST RESULTS.

Method	Bimanual Insertion			Put Box In Cabinet			Put Box In Bin			Put Box In Bin*		
	Top	Front	Both	Top	Left	Both	Top	Left	Both	Top	Left	Both
Diffusion Policy [10]	42%	44%	50%	16%	18%	26%	80%	64%	84%	38%	30%	44%
MAE-Diffusion	48%	50%	**54%**	42%	42%	<u>46%</u>	84%	78%	**92%**	52%	46%	**60%**
MAE-Select		<u>52%</u>			**50%**			<u>88%</u>			<u>58%</u>	

Method	Phone On Base			Pick Up Cup			Take Umbrella			Unplug Charger		
	Front	Wrist	Both	Front	Wrist	Both	Front	Wrist	Both	Top	Wrist	Both
Diffusion Policy [10]	82%	56%	78%	60%	40%	64%	58%	36%	54%	44%	30%	34%
MAE-Diffusion	86%	70%	<u>88%</u>	<u>68%</u>	66%	62%	56%	42%	**64%**	46%	34%	<u>52%</u>
MAE-Select		**92%**			**70%**			<u>60%</u>			**58%**	

Method	Put Eggplant To Bowl			Put Eggplant To Bowl*			Put Bitter Melon In Cabinet		
	Top	Wrist	Both	Top	Wrist	Both	Top	Left	Both
Diffusion Policy [10]	2/10	1/10	5/10	2/10	0/10	3/10	0/10	1/10	2/10
MAE-Diffusion	4/10	4/10	**7/10**	<u>4/10</u>	3/10	**6/10**	2/10	<u>4/10</u>	<u>4/10</u>
MAE-Select		<u>6/10</u>			**6/10**			**5/10**	

TABLE II

RESULTS OF PLUG-AND-PLAY EXPERIMENTS.

Method	Bimanual Insertion			Phone On Base		
	Top	Front	Both	Front	Wrist	Both
ACT [11]	14%	26%	34%	56%	50%	58%
MAE-ACT	28%	30%	**42%**	60%	58%	<u>66%</u>
MAE-Select		<u>36%</u>			**70%**	

TABLE III

ABLATION STUDIES ON MAE ENCODER AND DECODER UTILIZATION.

Method	Put Box In Cabinet			Phone On Base		
	Top	Left	Both	Front	Wrist	Both
MAE-Encoder	20%	28%	34%	76%	56%	80%
MAE-Diffusion	**42%**	**42%**	**46%**	**86%**	**70%**	**88%**

C. Single-camera control setup

We explore a single-camera control setup where the system is trained using multiple camera views but operates using a single camera during deployment. At each time chunk, the camera is positioned at one of the training viewpoints. This setup is particularly practical for scenarios where multiple cameras can be leveraged during training, while the robot must function with only one camera in real-world applications. Unlike MV-MWM [18], our approach allows the camera to move across different viewpoints during operation.

D. Results

We compare the performance of MAE-Select with two other methods: Diffusion Policy [10] and a variant of Diffusion Policy that incorporates MAE, referred to as MAE-Diffusion, as described in Sec.III-B, across different types of tasks and viewpoints. In the case of Diffusion Policy, the training and testing viewpoints are identical. For MAE-Diffusion, however, all available viewpoints are utilized during training.

As demonstrated in Tab.I, MAE-Select consistently outperforms other fixed single-camera setups in both simulation and real-world experiments. For example, in the *Put Box In Cabinet* task, MAE-Select improves performance by 8% compared to the best fixed single-camera method and by 32% compared to previous work. Its advantage lies in its ability to intelligently select the most informative viewpoints, which allows the system to make the most of limited visual input, resulting in optimized task completion.

Furthermore, an interesting pattern emerges in some tasks: for certain methods, the performance with a single viewpoint can surpass that of a multi-camera setup. For example, in the *Unplug Charger* task with Diffusion Policy, using only the top view (44%) outperforms using both views (34%). This counterintuitive result may stem from the added complexity of processing multiple cameras, which can introduce noise or misalignment issues, complicating the learning process. By focusing on the optimal viewpoint, MAE-Select avoids these challenges, enabling more efficient and effective task execution. Consequently, MAE-Select remains highly competitive when compared to multi-camera setups, even outperforming them in several tasks.

E. Plug-and-play

Our method emphasizes flexibility in viewpoint selection, making it independent of the specific action decoder used. To highlight this versatility, we also evaluate our approach in combination with ACT [11], an alternative action decoder. The results in Tab.II show that our viewpoint selection method can be seamlessly integrated with different action decoders, further showcasing its adaptability and plug-and-play capability in various system architectures.

Manuscript 2428 submitted to 2025 IEEE International Conference on Robotics and Automation (ICRA). Received September 15, 2024.

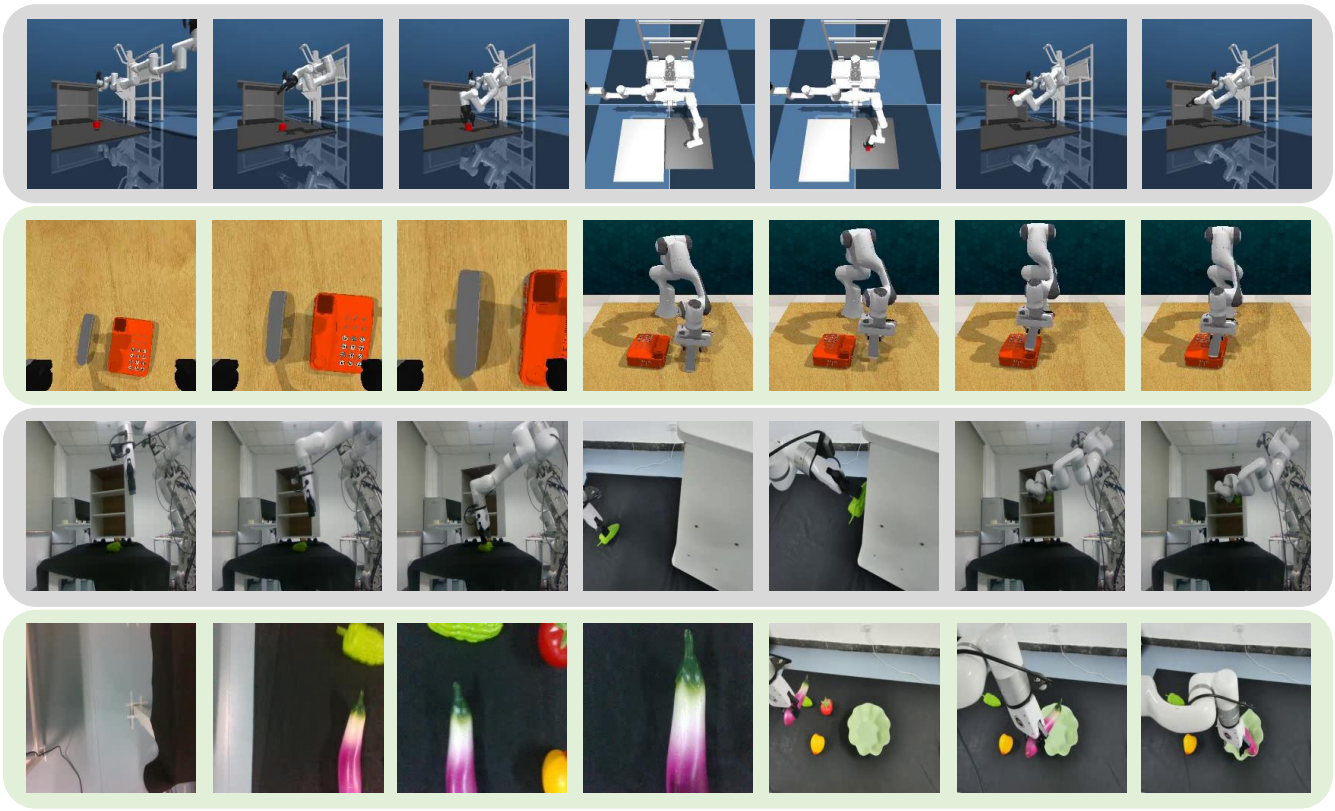

Fig. 3. Visualization of the selected viewpoints in our experiments, showcasing both simulation and real-world environments. Each row represents the whole procedure of a specific task, indicating the necessity of selecting different viewpoints throughout the task.

F. Ablation study

We conduct an ablation study to isolate the contributions of fully utilizing the entire encoder-decoder structure of the masked autoencoder in our approach. Specifically, we compare the performance of our method, which leverages both the encoder and decoder components, against a version that uses only the encoder of the masked autoencoder [37], [18]. The results in Tab.III demonstrate that utilizing the full encoder-decoder structure significantly improves performance, particularly in scenarios that require nuanced visual understanding from partial or occluded viewpoints. This highlights the value of the decoder in refining the system's ability to interpret and act based on incomplete or masked information, contributing to better generalization and adaptability in both simulated and real-world tasks.

G. Visualization

To better understand the behavior of the system under the single-camera control setup, we provide a comprehensive set of visualizations from deployment phases. As shown in Fig.3, MAE-Select selects optimal viewpoints based on contextual information, demonstrating its capacity for intelligent decision-making. In particular, the visualizations highlight how MAE-Select prioritizes critical areas of interest while disregarding less relevant regions.

V. CONCLUSIONS AND FUTURE WORK

In this work, we present MAE-Select, a novel framework that optimizes viewpoints for single-camera systems in robotic manipulation. By fully leveraging pre-trained representations from multi-view masked autoencoders and dynamically selecting the next most informative viewpoints at each time chunk without manual annotations, MAE-Select significantly enhances the efficiency of robotic manipulation, addressing the limitations of both multi-camera and single-view setups. Our experiments demonstrate that this plug-and-play mechanism effectively improves performance, and even surpasses multi-camera systems in certain cases. Despite its effectiveness, one major limitation of MAE-Select is that it optimizes over discrete viewpoints rather than continuous ones, which reduces the system's flexibility in dynamic environments. Future improvements could involve the integration of techniques like Neural Radiance Fields (NeRF) [43] or 3D Gaussian processes [44], enabling continuous viewpoint optimization.

REFERENCES

[1] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[2] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

[3] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*. PMLR, 2022, pp. 158–168.

[4] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," *arXiv preprint arXiv:2109.13396*, 2021.

[5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[7] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 750–16 761.

[8] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[10] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[11] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[12] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4788–4795.

[13] M. Reuss, M. Li, X. Jia, and R. Lioutikov, "Goal-conditioned imitation learning using score-based diffusion policies," *arXiv preprint arXiv:2304.02532*, 2023.

[14] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," *arXiv preprint arXiv:2112.01511*, 2021.

[15] Y. Qin, H. Su, and X. Wang, "From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 873–10 881, 2022.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[17] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[18] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 613–30 632.

[19] Y. Avigal, L. Berscheid, T. Asfour, T. Kröger, and K. Goldberg, "Speedfolding: Learning efficient bimanual folding of garments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1–8.

[20] S. P. Arunachalam, I. Güzey, S. Chintala, and L. Pinto, "Holo-dex: Teaching dexterity with immersive mixed reality," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5962–5969.

[21] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in neural information processing systems*, vol. 32, 2019.

[22] Y. Duan, M. Andrychowicz, B. Stadie, O. Jonathan Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[23] E. Johns, "Coarse-to-fine imitation learning: Robot manipulation from a single demonstration," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 4613–4619.

[24] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6401–6408.

[25] P. Florence, L. Manuelli, and R. Tedrake, "Self-supervised correspondence in visuomotor policy learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 492–499, 2019.

[26] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3758–3765.

[27] B. Wen, W. Lian, K. Bekris, and S. Schaal, "You only demonstrate once: Category-level manipulation from single visual demonstration," *arXiv preprint arXiv:2201.12716*, 2022.

[28] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 5628–5635.

[29] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024.

[30] B. Chen, P. Abbeel, and D. Pathak, "Unsupervised learning of visual 3d keypoints for control," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1539–1549.

[31] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[32] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[33] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.

[34] X. Li, W. Wang, L. Yang, and J. Yang, "Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality," *arXiv preprint arXiv:2205.10063*, 2022.

[35] X. Zhang, Y. Tian, W. Huang, Q. Ye, Q. Dai, L. Xie, and Q. Tian, "Hivit: Hierarchical vision transformer meets masked image modeling," *arXiv preprint arXiv:2205.14949*, 2022.

[36] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[37] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked world models for visual control," in *Conference on Robot Learning*. PMLR, 2023, pp. 1332–1344.

[38] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[39] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[40] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in psychology*. Elsevier, 1997, vol. 121, pp. 471–495.

[41] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, 2020.

[42] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.

[43] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[44] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.