

Robust Imitation Learning for Mobile Manipulator Focusing on Task-Related Viewpoints and Regions

Yutaro Ishida¹, Yuki Noguchi¹, Takayuki Kanai¹, Kazuhiro Shintani¹ and Hiroshi Bito¹

Abstract—We study how to generalize the visuomotor policy of a mobile manipulator from the perspective of visual observations. The mobile manipulator is prone to occlusion owing to its own body when only a single viewpoint is employed and a significant domain shift when deployed in diverse situations. However, to the best of the authors’ knowledge, no study has been able to solve occlusion and domain shift simultaneously and propose a robust policy. In this paper, we propose a robust imitation learning method for mobile manipulators that focuses on task-related viewpoints and their spatial regions when observing multiple viewpoints. The multiple viewpoint policy includes attention mechanism, which is learned with an augmented dataset, and brings optimal viewpoints and robust visual embedding against occlusion and domain shift. Comparison of our results for different tasks and environments with those of previous studies revealed that our proposed method improves the success rate by up to 29.3 points. We also conduct ablation studies using our proposed method. Learning task-related viewpoints from the multiple viewpoints dataset increases robustness to occlusion than using a uniquely defined viewpoint. Focusing on task-related regions contributes to up to a 33.3-point improvement in the success rate against domain shift.

I. INTRODUCTION

In recent years, mobile manipulators (MMs) have entered the spotlight owing to their ability to perform various tasks in various locations, which can be attributed to their high mobility [1]–[4]. Extensive research and development have been conducted to explore their societal implementation, including recognition [5]–[7], control [8], autonomy [9], simulation [10] and robotics competitions [11], [12]. Researchers are exploring robot learning techniques such as imitation learning and reinforcement learning to develop learned policies that can achieve more advanced tasks than traditional hand-crafted programming [13]–[16].

In robot learning, a method called visuomotor policy is commonly used to predict the next action based on visual observation. However, there are two problems that are typically encountered when using the policy on MMs.

First, MMs are prone to occlusion in visual observation when only a single internal viewpoint is used. For instance, in the pick task shown in Fig. 1 left, the first-person viewpoint can easily be occluded by the body of the MM. In contrast, in the place task shown in Fig. 1 right, the in-hand viewpoint can easily be occluded by the grasped object. To address this, the policy needs to be able to handle occlusion by focusing on task-related viewpoints from multiple internal viewpoints

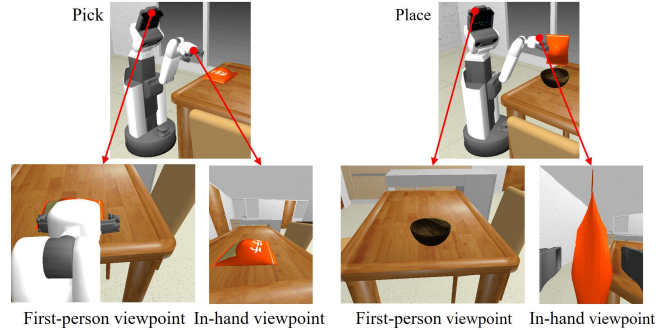


Fig. 1. Example of occlusion of internal viewpoints on the mobile manipulators. Left: first-person viewpoint is occluded by the body of the MM in pick task. Right: in-hand viewpoint is occluded by the grasped object in place task.



Fig. 2. Example of visual observation domain shift. Left: environments for training the policy. Middle: distractor objects cause the minor change. Right: unknown furniture causes the major change.

depending on the situation: using in-hand viewpoint for pick task (Fig. 1 left) and first-person viewpoint for place task (Fig. 1 right).

Second, compared with a manipulator without a mobile base, MMs are more prone to significant domain shifts in visual observation owing to their varied usage [14], [17]. For example, in the pick-red-can task, training environments (Fig. 2 left) are called in-distribution (ID), while other environments with distractor objects (Fig. 2 middle) or unfamiliar furniture (Fig. 2 right) are called out-of-distribution (OOD). This domain shift is particularly noticeable with RGB images but might be suppressed with other modalities, such as depth images. However, the semantics provided by RGB images are crucial for handling task-specific objects among various ones. Thus, the policy should mainly focus on task-related regions in visual observations and avoid overreacting to domain shifts in non-task-related regions.

In the previous study [18], the policy focuses on task-related viewpoints by uniquely defining from multiple viewpoints to improve robustness to domain shift. The viewpoints

¹Toyota Motor Corporation, 1 Toyota-Cho, Toyota City, Aichi 471-8571, Japan {yutaro-ishida, yuki.noguchi@ad, takayuki.kanai, kazuhiro.shintani, hiroshi.bitto}@mail.toyota.co.jp

that underwent minimal changes under domain shift were considered important for the task, and a variational information bottleneck (VIB) was added to the back end of the image encoder for the other viewpoints to filter the information. The study concluded that an in-hand viewpoint is more robust than third-person viewpoints against changes in robot position, distractor objects, and background. However, the study did not aim to improve robustness to occlusion, and a uniquely defined viewpoint is not necessarily suitable for all occluded situations as mentioned above. Additionally, just focusing on one unique defined viewpoint is not robust enough to address domain shift. The policy requires a mechanism that can focus on not only task-related viewpoints but also their spatial regions.

In another previous study [14], [19] used a large dataset to enable learning of task-related regions in the policy for improving robustness to domain shift. A large dataset is collected through long hours of teleoperated expert demonstration and photo-realistic augmentation using a diffusion model. By learning the dataset with a policy that includes spatial attention, task-related regions can be focused on. However, the study was not conducted in a setting with multiple viewpoints, and the performance might decrease for a single internal viewpoint, which can be easily occluded in MMs. Additionally, although augmentation by the diffusion model is photo-realistic, it has high time and computation demands, which increases in proportion to the number of viewpoints.

In this paper, we propose a robust imitation learning method for MMs that focuses on task-related viewpoints and their spatial regions when observing multiple viewpoints. First, we propose an attention mechanism for the multiple viewpoints and their spatial regions. Second, to facilitate the learning of attention mechanism, we propose a method of fast and low computational resource augmentation using fractal texture for non-task-related regions. This approach allows the policy to focus on task-related regions across multiple viewpoints and improves the robustness of occlusion and domain shift. Comparison of our results for different tasks and environments with those of previous studies revealed that our proposed method improves the success rate by up to 29.3 points. We also conduct ablation studies using our proposed method. Learning task-related viewpoints from the multiple viewpoints dataset increases robustness to occlusion than using a uniquely defined viewpoint. Focusing on task-related regions contributes to up to a 33.3-point improvement in the success rate against domain shift.

II. RELATED WORK

A. Mobile Manipulator

Mobile manipulators have the ability to move and operate in various locations owing to their mobile base. Moreover, the mobile base allows them to change their posture while they are performing tasks. Therefore, compared to a manipulator that does not have a mobile base, MMs have the potential to perform a variety of tasks [1], [8].

The intelligence of MMs has been realized with hand-craft designed algorithms that can be roughly categorized into perception, decision-making, and control algorithms [7]. However, the implementation by skilled engineers requires a considerable amount of time [12]. Additionally, even with the time investment, it is not possible to implement solutions for modeling difficult algorithms by humans such as a manipulator cooking something [15]. Therefore, an MM is expected to learn behaviors through mimicking and trial-and-error by itself using robot learning [13], [14], [16], [17].

B. Robot Learning

Robots learn behavior through reinforcement learning, which has been studied extensively [20], [21]. In the learning process, the robot is forced to actuate on random or unstable policies to collect data. This approach carries the risk of malfunction, and its application has been limited to problems where the environment can be automatically reset after each episode. In recent years, robots have been expected to learn behaviors by imitation learning [14]–[16], [18], [19]. However, as mentioned in Section I, to the best of the authors’ knowledge, there are no study on robust policies for MMs, which are prone to occlusion and domain shift.

Following the recent success of foundation models in language and image fields, robot learning has also been improved in terms of architecture and data volume. In terms of architecture, there are many studies on transformer-based policies [14], [15], [17]. There are also studies on conditional diffusion models that generate robot trajectories [22]–[24]. In terms of data volume, attempts have been made to collect expert demonstrations through long hours of teleoperations [14]. However, it is not practical to collect expert demonstrations of MMs for all situations with teleoperations. Therefore, there is a need to improve data efficiency by reducing the domain shift which non-task related things by, for example, reducing the dimension of embedding using spatial attention [25], which is also used in the foundation model.

III. PRELIMINARIES

Definitions of Terms: We consider the scenario where an MM solves a given task by interacting in the environment. Specifically, we consider that MMs with parallel grippers manipulate objects as shown in Fig. 1. In this setting, we define *task-related regions* as the spatial region of objects related to the task in the camera image. The other regions are defined as *non-task-related regions*. For example, in the pick-red-can task shown in Fig. 2, task-related regions are the pixels corresponding to the red can, and non-task-related regions are any other pixels.

Imitation Learning: We consider imitation learning in which a robot mimics an expert demonstration and acquires a policy using machine learning. The expert demonstration is a dataset $\mathcal{D} = \{(o_{\text{expert}_t}^{(n)}, a_{\text{expert}_t}^{(n)})_{t=0}^{T^{(n)}}\}_{n=0}^N$ of tasks successfully solved by teleoperation. o_{expert} and a_{expert} are the robot’s observations and actions acquired during teleoperation, T is the timestep of each episode, and N is

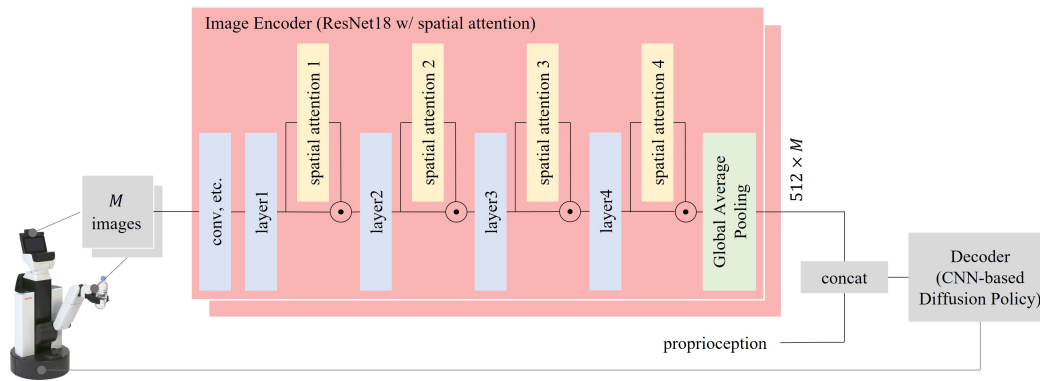


Fig. 3. Attention mechanism for multiple viewpoints and their spatial regions. By weighting the features with spatial attention, the information of task-related viewpoints and their spatial regions are extracted in image encoders from multiple visual observations. Since spatial attention is the learnable parameter, our method can learn task-related viewpoints from dataset instead of uniquely defined by hand-craft.

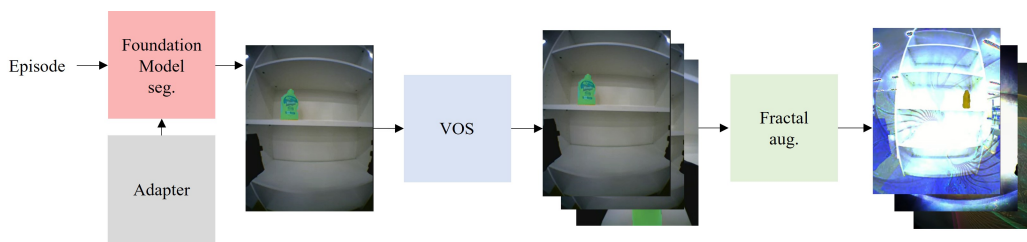


Fig. 4. Processing steps of fast and low computational resource augmentation using fractal texture. By detecting and tracking task-related regions, non-task-related regions are augmented with fractal textures. The augmentation facilitates the learning of attention mechanism that focuses strongly on task-related regions which less changed, rather than non-task-related regions that are changed greater with fractal texture.

the number of episodes. The policy $\pi(a|o)$ is trained by supervised learning using dataset \mathcal{D} to minimize the error between predicted action a_{pred} and a_{expert} when o_{expert} is given as the input. At test time, the robot works by continuing the process of inputting the current observation o and predicting action a using policy $\pi(a|o)$.

Observations and Actions: We focus on MMs with internal multiple sensors as shown in Fig. 1. As a type of observation, we consider in-hand viewpoint o_h , first-person viewpoint o_f and proprioception o_p . The action a is the relative movement of the end-effector frame. The low-level controller receives a from the policy and solves inverse kinematics to convert it into a joint position command value.

IV. PROPOSED METHOD

We propose a robust imitation learning method for MMs that focuses on task-related viewpoints and their spatial regions when observing multiple viewpoints. First, we propose an attention mechanism for multiple viewpoints and their spatial regions. Second, to facilitate the learning of spatial attention, we propose a method of fast and low computational resource augmentation using fractal texture for non-task-related regions. These methods enable the policy to focus on task-related regions across multiple viewpoints and improve the robustness of occlusion and domain shift.

A. Attention Mechanism for Multiple Viewpoints and Their Spatial Regions for Imitation Learning Policy

Our proposed method modifies a common architecture of imitation learning with visual observation, which combines a network of encoders and decoders. Here we illustrate our proposed method using an example of an encoder and decoder policy, Diffusion Policy (DP) [24] (Fig. 3). As a first modification, our proposed method focuses on task-related regions to improve the robustness of domain shift. Following a previous study [14] that conducted a single viewpoint, our proposed method incorporates spatial attention inside the image encoder as shown in Fig. 3. The spatial attention predicts weights $w(b, h, w, 1)$ using multiple convolution layers for a feature $f(b, h, w, c)$ on a convolutional neural network, expands it to channel dimension, and finally takes the inner product of $f(b, h, w, c)$ and $w(b, h, w, c)$. The features are weighted to task-related regions. As a second modification, our proposed method handles multiple viewpoints to improve the robustness against occlusion. Our proposed method includes per-viewpoint image encoders instead of a shared encoder to focus spatial regions in each viewpoint as shown in Fig. 3. The image encoders for M viewpoints have parameters corresponding to each viewpoint.

Our proposed method focuses on the optimal viewpoints of a task by keeping only the information of task-related viewpoints from multiple viewpoints using the attention mechanism. Contrary to a previous study [18] that uniquely

TABLE I
HYPERPARAMETERS FOR DP [24].

Ctrl	To	Ta	Tp	ImgRes	CropRes	#D-Params	#V-Params	Lr	WDecay	D-Inters Train	D-Inters Eval
Pos	2	8	16	2x320x240	2x288x216	67 millions	22 millions	1e-4	1e-6	100	8

defined optimal single viewpoints, our method can learn task-related viewpoints from a dataset. Thus, it can be applied for tasks that are performed under various conditions. Additionally, our proposed method acquires robust visual embedding against domain shift by keeping only the information of task-relevant spatial regions using the attention mechanism. The embedding is robust against large distribution shifts in visual observation caused by distractor objects or unknown furniture as shown in Fig. 2.

B. Fast and Low Computational Resource Augmentation using Fractal Texture

To facilitate the learning of attention mechanism, we propose a method of augmenting non-task related regions with fractal texture (Fig. 4). The attention mechanism focuses strongly on task-related regions that have undergone minimal changes rather than non-task-related regions that have undergone considerable changes after application of fractal texture. Contrary to the previous study [19], our proposed method has minimal time and computational resources, making it suitable for augmenting multiple viewpoints with a larger volume of images compared to a single viewpoint.

First, our proposed method detects task-related regions in visual observations o_h and o_f at $t = 0$ in dataset \mathcal{D} as shown Fig. 4 left. Foundation models adapted to robotic training data, such as a combination of FastSAM [26] and CLIP-Adapter [27], use text prompts to detect task-specific objects. (Note that FastSAM is adapted to in-distribution training data and requires hyperparameters related to segmentation granularity to be tuned according to the input, which make it less effective in detecting with out-of-distribution data during execution.) Second, the task-related regions are tracked from $t = 0$ to $t = T(n)$ as shown Fig. 4 middle. XMem [28] is the video object segmentation (VOS) method used for tracking. These steps generate non-task-related region masks, which is complement of task-related regions, for all times during all the episodes. Finally, augmentation is applied to non-task related regions using PixMix [29] as shown Fig. 4 right. Objects with patterns ranging from simple contours, such as a box, to complex contours, such as fern leaves, appear in non-task-related regions. Fractals can be controlled by the complexity of their patterns with parameters and have a structure in which a part is self-similar to the whole, for example, a pattern similar to the original appears even after repeated enlargements. Therefore, the augmented dataset simulates artificial and natural objects by various patterns of fractals and is as effective as a large dataset with various backgrounds collected through long hours of teleoperation.



Fig. 5. Overview of the pick-bottle-from-shelf task. Figures are lined in time-step order from left to right. Left: the MM started with o_h and o_f facing the bottle placed on the shelf. Middle: the MM moved the mobile base and arm to reach the bottle. Right: the MM picked up the bottle from the shelf.



Fig. 6. List of pick-bottle-from-shelf task environments. Left: In distribution (ID). Middle: Out of distribution due to the distractor objects (OOD-distractor). Right: Out of distribution due to the unknown shelf (OOD-shelf).

V. EXPERIMENTS

To validate the effectiveness of our proposed method, we performed experiments by varying the task and environment. In the following section, we describe the experimental settings, compare our results with those of previous studies, and describe ablation studies to answer the following three questions.

- 1) What is the effect of the multiple viewpoints?
- 2) How can MM focus on task-related viewpoints?
- 3) How can MM focus on task-related regions?

A. Settings: Tasks and Environments

pick-bottle-from-shelf: Figure 5 shows the sequence of the pick-bottle-from-shelf task in time-step order. The MM started by observing the bottle placed on the shelf with o_h or o_f (Fig. 5 left), reached the bottle (Fig. 5 middle), and picked up the bottle from the shelf (Fig. 5 right). If the MM could lift the bottle off the shelf board and carry it out of the shelf, it was successful in completing the task. In this task, o_h had no occlusion to observe the bottle and o_f had occlusion due to the body of the MM (Fig. 9 right). Thus, o_h had more information on the visuomotor policy.

Three environmental variations are shown in Fig. 6. The ID environment in Fig. 6 left was the same as when the expert demonstrations were collected. The OOD-distractor environment in Fig. 6 middle had the same shelf as when the expert demonstration was collected, but there were distractor objects around the bottle. The OOD-shelf environment on Fig. 6 right had an unknown shelf. The expert demonstrations

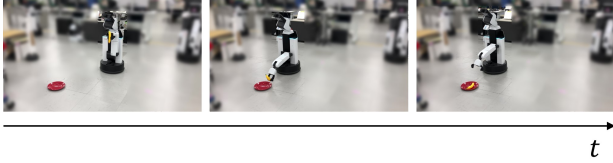


Fig. 7. Overview of the place-banana-on-plate task. Figures are lined in time-step order from left to right. Left: the MM started by grasping the banana with o_f facing the plate on the floor. Middle: the MM moved the mobile base and arm to reach the plate. Right: the MM placed and left the banana on the plate.

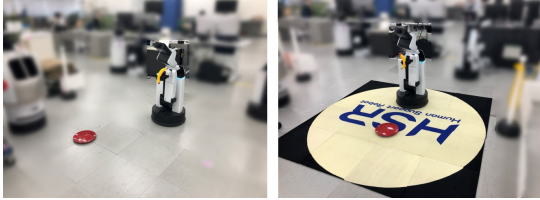


Fig. 8. List of place-banana-on-plate task environments. Left: In distribution (ID). Right: Out of distribution due to the unknown floor texture (OOD-shelf).

were collected 50 times in the ID environment, and each environment was tested 15 times while varying the positions of the MM, the bottle, and distractor objects.

place-banana-on-plate: Figure 7 shows the sequence of the place-banana-on-plate task in time-step order. The MM started by observing the plate placed on the floor with o_f (Fig. 7 left), reached the plate (Fig. 7 middle), and placed the banana on the plate (Fig. 7 right). If the MM could place the banana on the plate without touching the floor, it was successful in completing the task. At the beginning of the time-step, o_f had no occlusion to observe the plate, while o_h only observed the banana and the background (Fig. 10 left). However, in the late stage of the time-step, o_f experienced occlusion due to the body of the MM, and o_h could observe the plate (Fig. 10 right). Thus, both o_f and o_h had the information on the visuomotor policy such as the position of the plate.

Two environmental variations are shown in Fig 8. The ID environment in Fig. 8 left was the same as when the expert demonstrations were collected. The OOD-floor environment on Fig. 8 right had an unknown floor texture. The expert demonstrations were collected 50 times in the ID environment, and each environment was tested 15 times while varying the positions of the MM and the plate.

B. Settings: Implementations

The following three methods were implemented for comparison: (1) Previous study: Original DP [24]; (2) previous study: DP with VIB on each back end of image encoders and augmentation (Aug) [18], and (3) proposed method: DP with attention mechanism (AM) and augmentation. The DP hyperparameters are shown in Table I. The architecture of VIB and spatial attention are shown in Table II and Table III respectively. β used in the VIB loss function was set to

TABLE II
ARCHITECTURE OF VIB

	Layer Description	Output Tensor Dim.
#0	Input feature	512
	Encoder	
#1	Linear + ReLU	256
#2	Linear + ReLU	256
	μ layer	
#3-1	Linear	128
	σ layer	
#3-2	Linear + Softplus	128
	Latent layer	
#4	$\mu + \sqrt{\sigma} \times randn$	128
	Decoder	
#5	Linear	512

TABLE III
ARCHITECTURE OF SPATIAL ATTENTION

	Layer Description	Output Tensor Dim.
#0	Input feature	$c \times h \times w$
#1	Conv2d + ReLU	$c \times h \times w$
#2	Conv2d + Sigmoid	$1 \times h \times w$
#3	expand (Channel Dim.)	$c \times h \times w$

1.0. We used the Toyota Human Support Robot (HSR) [8] as the MM.

C. Overall Results

Comparison of the results of previous studies and our proposed method in the pick-bottle-from-shelf task and place-banana-on-plate tasks are shown in Table IV and Table V respectively. As described in Section V-B, the results of two previous studies were compared with the results of our proposed method. All methods had three observations o_h, o_f, o_p . Following the previous studies, VIB was applied for either or both of encoded o_h (z_h) and o_f (z_f).

Compared to the original DP, our proposed method improved the success rate by 6.6 points in the pick-bottle-from-shelf ID environment (Table IV). No improvement was observed in the place-banana-on-plate ID environment (Table V). A large success rate difference appeared in OOD environments. Our proposed method demonstrated improvements of 33.3 points in pick-bottle-from-shelf OOD-distractor, 100.0 points in pick-bottle-from-shelf OOD-shelf, and 13.4 points in place-banana-on-plate OOD-floor respectively (Table IV and V). Even after including degradation, which was limited to one failure in 15 tests of one environment, our proposed method improved the average success rate for the five environments by 29.3 points. The improvement in the 4/5 environment was due to the attention mechanism and augmentation being robust to domain shift of non-task related regions, as described in Section IV.

Compared to the original DP, VIB exhibited the same trend reported in a previous study and improved the success rate in the pick-bottle-from-shelf task in the order of

TABLE IV
OVERALL RESULTS OF THE PICK-BOTTLE-FROM-SHELF TASK

method	ID	Success Rate [%]	
		OOD (distracter)	OOD (shelf)
Original [24]	86.7	46.7	0.0
VIB(z_h) + Aug [18]	0.0	6.7	6.7
VIB(z_f) + Aug [18]	93.3	73.3	86.7
VIB(z_h) + VIB(z_f) + Aug [18]	86.7	40.0	60.0
Proposal: AM + Aug	93.3	80.0	100.0

TABLE V
OVERALL RESULTS OF THE PLACE-BANANA-ON-PLATE TASK

method	Success Rate [%]	
	ID	OOD (floor)
Original [24]	100.0	73.3
VIB(z_h) + Aug [18]	60.0	33.3
VIB(z_f) + Aug [18]	66.7	20.0
VIB(z_h) + VIB(z_f) + Aug [18]	13.3	0.0
Proposal: AM + Aug	93.3	86.7

VIB(z_f), VIB(z_h, z_f) and VIB(z_h) (Table IV). However, VIB exhibited a different trend, and no improvement was observed in the place-banana-on-plate task (Table V). The reason for the different trend was that VIB uniquely defined a viewpoint as mentioned in Section I and could not handle cases of changing task-related viewpoints during execution as mentioned in Section V-A.

Compared to VIB(z_f), which was considered the best in previous studies, our proposed method achieved an equal or higher success rate in all tasks and environments. Our proposed method demonstrated improvements of 6.7 points in pick-bottle-from-shelf OOD-distracter, 13.3 points in pick-bottle-from-shelf OOD-shelf, 26.6 points in place-banana-on-plate ID, and 66.7 points in place-banana-on-plate OOD-

TABLE VI
ABLATION STUDY: EFFECT OF MULTIPLE VIEWPOINTS
IN THE PICK-BOTTLE-FROM-SHELF TASK

observation	ID	Success Rate [%]	
		OOD (distracter)	OOD (shelf)
o_h, o_p	100.0	80.0	80.0
o_f, o_p	20.0	6.7	6.7
Proposal: o_h, o_f, o_p	93.3	80.0	100.0

TABLE VII
ABLATION STUDY: EFFECT OF MULTIPLE VIEWPOINTS
IN THE PLACE-BANANA-ON-PLATE TASK

observation	Success Rate [%]	
	ID	OOD (floor)
o_h, o_p	26.7	26.7
o_f, o_p	86.7	66.7
Proposal: o_h, o_f, o_p	93.3	86.7



Fig. 9. Visual observations in the pick-bottle-from-shelf task. Figures are lined in time-step order from left to right.

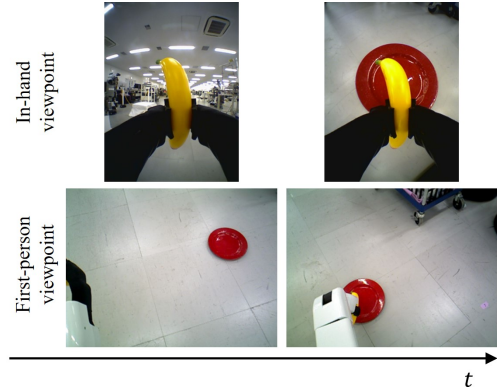


Fig. 10. Visual observations in the place-banana-on-plate task. Figures are lined in time-step order from left to right.

floor respectively (Table IV and V). Our proposed method improved the average success rate of the five environments by 22.7 points. These results were achieved because our proposed method can learn task-related viewpoints from datasets in various situations and could focus on task-related regions as mentioned in Section IV (Detailed in Section V-E and V-F).

D. What is the effect of the multiple viewpoints?

To clarify the effect of multiple viewpoints, the ablation study of viewpoints in the pick-bottle-from-shelf and place-banana-on-plate tasks are shown in Table VI and Table VII respectively. All methods were based on our proposed method described in section V-B, and three types of observations are compared: only in-hand viewpoint o_h, o_p , only first-person viewpoint o_f, o_p , and multiple viewpoints o_h, o_f, o_p .

In the pick-bottle-from-shelf task, the in-hand viewpoint achieved a higher success rate than the first-person viewpoint (Table VI). On the other hand, in the place-banana-on-plate task, the in-hand viewpoint achieved a lower success rate than the first-person viewpoint (Table VII). As described in Section V-A, each task had different occluded viewpoints, and use of a single viewpoint did not provide robustness against occlusion (Fig. 9 and 10). Table VIII shows the

TABLE VIII
STRATIFIED ANALYSES OF THE INITIAL DISTANCE BETWEEN
THE MM AND THE PLATE IN THE PLACE-BANANA-ON-PLATE TASK

observation	Success Rate [%]	
	< 0.7 [m]	≥ 0.7 [m]
o_h, o_p	44.4	19.0
o_f, o_p	88.9	71.4
Proposal: o_h, o_f, o_p	100.0	85.7

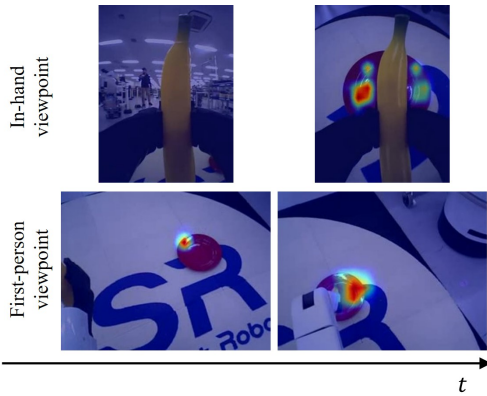


Fig. 11. Visualize pixel attribution on the prediction of our proposed method in the place-banana-on-plate task. Figures are lined in time-step order from left to right. The method was able to focus on the plate, which changes visibility moment by moment.

stratified analyses of the initial distance between the MM and the plate in the place-banana-on-plate task. The difference between the in-hand viewpoint and first-person viewpoint was 44.5 points for short distances and 52.4 points for long distances. For short-distance cases, the MM might observe the plate when the in-hand viewpoint is pointed at the floor in a guesswork manner. On the contrary, the MM cannot observe the plate for long-distance cases. Thus, the number of failures increase.

For both the pick-bottle-from-shelf and place-banana-on-plate tasks, the use of multiple viewpoints resulted in higher success rates than the use of in-hand viewpoint or first-person viewpoint (Table VI and VII). Thus, multiple viewpoints provide robust visual observation even in cases where occlusion occurs, such as when the internal viewpoints of the MM are used.

E. How can MM focus on task-related viewpoint?

As mentioned in Section V-C, compared to VIB(z_f), which was considered the best in previous studies, our proposed method achieved an equal or high success rate in all tasks and environments. Figure 11 shows qualitative analysis with visualize the pixel attribution on the prediction of our proposed method using GradCAM [30] in the place-banana-on-plate task where a difference of success rate was particularly large. At the beginning of the time-step, the method focused on the plate in o_f , as shown in Fig. 11 left. In the late stage of the time-step, the method focused on the plate in both o_h and o_f , as shown in Fig. 11 right.

TABLE IX
ABLATION STUDY OF HOW TO FOCUS TASK-RELATED REGIONS

method	ID	Success Rate [%]	
		OOD (distracter)	OOD (shelf)
AM	100.0	73.3	0.0
Aug	86.7	66.7	66.7
Proposal: AM + Aug	93.3	80.0	100.0

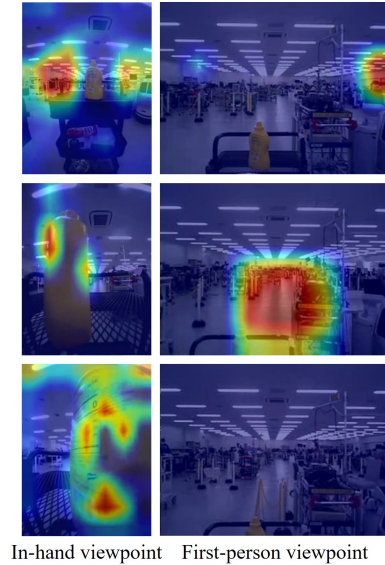


Fig. 12. Visualize pixel attribution on the prediction of following methods in pick-bottle-from-shelf task. Top: only attention mechanism. Middle: only augmentation. Bottom: both attention mechanism and augmentation.

Compared to VIB which uniquely defined viewpoints, our proposed method could focus on optimal viewpoints based on the situation and adapt well to various tasks.

F. How can MM focus on task-related regions?

To clarify how the policy focused on task-related regions, the ablation study in the pick-bottle-from-shelf task is shown in Table IX. All methods are based on our proposed method described in section V-B, and three types are compared: only attention mechanism is used, only augmentation is used, and both attention mechanism and augmentation are used. We visualized the pixel attribution on the prediction using GradCAM.

In the case where only attention mechanism was used, the policy was unable to focus on task-related regions (bottle) as shown in Fig. 12 (top), and decreased the average success rate of the three environments by 33.3 points. In the case of only augmentation, the policy was able to focus on task-related regions with o_h as shown in Fig. 12 (left of middle). However, the policy was unable to focus on task-related regions with o_f as shown in Fig. 12 (right of middle). The attribution was frozen during task execution as shown in Fig. 12 (right of middle), which showed that the policy could not learn task-related regions correctly. The policy decreased the average success rate of the three environments by 17 points.

In the case of both attention mechanism and augmentation, the policy was able to focus on task-related regions as shown in Fig. 12 (bottom), and resulted in the best success rate under domain shift. Thus, the result indicates the importance of ensuring a mechanism that focuses on task-related regions with attention mechanism and facilitates the learning of that with augmentation.

VI. CONCLUSIONS

We propose a robust imitation learning method for MMs that focuses on task-related viewpoints and spatial regions when observing multiple viewpoints. The multiple viewpoint policy includes attention mechanism, which is learned with an augmented dataset, and brings optimal viewpoints and robust visual embedding against occlusion and domain shift.

In the future, the model should be validated in dynamic and more complex environments. For this purpose, the model should be extended to use other observation modalities for considering cases in which changes cannot be observed visually. Its robustness to changes in task-related regions, such as changes in bottle color, should be enhanced. Additionally, its robustness should extend beyond changes in observations to encompass changes in actions as well.

REFERENCES

- [1] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of the research platform of a domestic mobile manipulator utilized for international competition and field test," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 7675–7682.
- [2] "Robots/pr2 - ros wiki," accessed on 10 Feb 2024. [Online]. Available: <http://wiki.ros.org/Robots/PR2>
- [3] "Fetch & freight manual," accessed on 10 Feb 2024. [Online]. Available: <http://docs.fetchrobotics.com/>
- [4] "Tiago," accessed on 10 Feb 2024. [Online]. Available: <https://pal-robotics.com/robots/tiago/>
- [5] Y. Ishida and H. Tamukoh, "Semi-automatic dataset generation for object detection and recognition and its evaluation on domestic service robots," *Journal of Robotics and Mechatronics*, vol. 32, no. 1, pp. 245–253, 2020.
- [6] Y. Ishida, T. Morie, and H. Tamukoh, "A hardware intelligent processing accelerator for domestic service robots," *Advanced Robotics*, vol. 34, no. 14, pp. 947–957, 2020.
- [7] T. Ono, D. Kanaoka, T. Shiba, S. Tokuno, Y. Yano, A. Mizutani, I. Matsumoto, H. Amano, and H. Tamukoh, "Solution of world robot challenge 2020 partner robot challenge (real space)," *Advanced Robotics*, vol. 36, no. 17–18, pp. 870–889, 2022.
- [8] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of human support robot as the research platform of a domestic mobile manipulator," *ROBOMECH Journal*, vol. 6, no. 4, 2019.
- [9] T. Yamamoto, H. Yaguchi, S. Kato, and H. Okada, "Evaluation of impression difference of a domestic mobile manipulator with autonomous and/or remote control in fetch-and-carry tasks," *Advanced Robotics*, vol. 34, no. 20, pp. 1291–1308, 2020.
- [10] Y. Matsusaka, L. Contreras, H. Okada, Y. Iwanaga, and T. Yamamoto, "A continuous integration based simulation environment for home support robot and its application to robocup competition," in *2023 IEEE/SICE International Symposium on System Integration (SII)*, 2023, pp. 1–6.
- [11] "Robocup," accessed on 10 Feb 2024. [Online]. Available: <https://www.robocup.org/>
- [12] L. Contreras, T. Yamamoto, Y. Matsusaka, and H. Okada, "Towards general purpose service robots: World robot summit – partner robot challenge," *Advanced Robotics*, vol. 36, no. 17–18, pp. 812–824, 2022.
- [13] 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 *5th Annual Conference on Robot Learning*, 2021.
- [14] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," in *arXiv preprint arXiv:2212.06817*, 2022.
- [15] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *arXiv preprint arXiv:2304.13705*, 2023.
- [16] Z. Fu, T. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *arXiv preprint arXiv:2401.02117*, 2024.
- [17] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choro-manski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *arXiv preprint arXiv:2307.15818*, 2023.
- [18] K. Hsu, M. J. Kim, R. Rafailov, J. Wu, and C. Finn, "Vision-based manipulators need to also see from their hands," in *International Conference on Learning Representations*, 2022.
- [19] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, D. M. J. Peralta, B. Ichter, K. Hausman, and F. Xia, "Scaling robot learning with semantically imagined experience," in *arXiv preprint arXiv:2302.11550*, 2023.
- [20] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "Mt-opt: Continuous multi-task robotic reinforcement learning at scale," in *arXiv preprint arXiv:2104.08212*, 2021.
- [21] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," in *arXiv preprint arXiv:1812.00568*, 2018.
- [22] M. Janner, Y. Du, J. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *International Conference on Machine Learning*, 2022.
- [23] A. Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision making?" in *The Eleventh International Conference on Learning Representations*, 2023.
- [24] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [25] M. S. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: What can 8 learned tokens do for images and videos?" in *arXiv preprint arXiv:2106.11297*, 2021.
- [26] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," in *arXiv preprint arXiv:2306.12156*, 2023.
- [27] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," in *arXiv preprint arXiv:2110.04544*, 2021.
- [28] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *European Conference on Computer Vision*, 2022.
- [29] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "Pixmix: Dreamlike pictures comprehensively improve safety measures," in *Computer Vision and Pattern Recognition*, 2022.
- [30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.