Active Vision Might Be All You Need: Exploring Active Vision in Bimanual Robotic Manipulation

Ian Chuang[*1,2] **Andrew Lee**[*2] **Dechen Gao**[2] **Mahdi Naddaf**[2] **Iman Soltani**[2]
[1] University of California Berkeley [2] University of California Davis

Abstract: Imitation learning has demonstrated significant potential in performing high-precision manipulation tasks using visual feedback from cameras. However, it is common practice in imitation learning for cameras to be fixed in place, resulting in issues like occlusion and limited field of view. Furthermore, cameras are often placed in broad, general locations, without an effective viewpoint specific to the robot's task. In this work, we investigate the utility of active vision (AV) for imitation learning and manipulation, in which, in addition to the manipulation policy, the robot learns an AV policy from human demonstrations to dynamically change the robot's camera viewpoint to obtain better information about its environment and the given task. We introduce AV-ALOHA, a new bimanual teleoperation robot system with AV, an extension of the ALOHA 2 robot system, incorporating an additional 7-DoF robot arm that only carries a stereo camera and is solely tasked with finding the best viewpoint. This camera streams stereo video to an operator wearing a virtual reality (VR) headset as feedback, allowing the operator to control the camera pose using head and body movements. The system provides an immersive teleoperation experience, with bimanual first-person control, enabling the operator to dynamically explore and search the scene and simultaneously interact with the environment. We conduct imitation learning experiments of our system both in real-world and in simulation, across a variety of tasks that emphasize viewpoint planning. Our results demonstrate the effectiveness of human-guided AV for imitation learning, showing significant improvements over fixed cameras in tasks with limited visibility. Project website: https://soltanilara.github.io/av-aloha/

Keywords: Active Vision, Bimanual manipulation, Imitation learning

1 Introduction

Recent advances in robot learning architectures [1, 2] along with the development of low-cost, open-source methods for easier robot data collection [1, 3], have led to an accelerated advancement for robot learning using imitation learning methods [4, 5, 6]. End-to-end imitation learning-based approaches offer a scalable and general solution to bimanual tasks that would be very challenging to implement using heuristic-based, task-specific methods. A key feature of these systems is that instead of relying on precise calibration and expensive sensors, these systems can achieve remarkable precision by instead relying on visual feedback from inexpensive cameras [1].

In most robotics implementations, it is common for cameras to be either fixed in place [1, 7, 8] or mounted eye-in-hand in combination with a tool like a gripper [3, 9]. These cameras are typically positioned in a task-agnostic manner, without considering specific visibility requirements of a given task. However, optimal camera placement is crucial to enable effective learning and execution. With an inadequate camera viewpoint, these models will struggle, especially in situations

[*]Equal contribution

CoRL 2024 Workshop on Whole-Body Control and Bimanual Manipulation (CoRL 2024 WCBM).

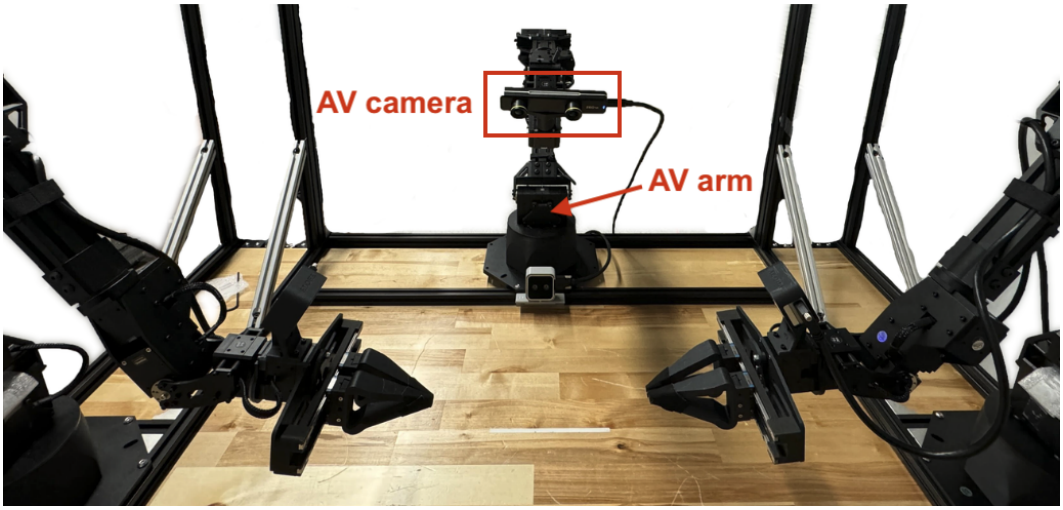

Figure 1: We introduce AV-ALOHA, a bimanual robot system with 7-DoF AV. In this system, a VR headset provides a live feed from the AV camera to the user. The movement of the VR headset controls the AV arm.

where the object being handled occludes the camera's view or when tasks require close-ups of small features [1]. Consider an assembly task, such as a peg-in-hole scenario where the location of the hole has limited visibility, like threading a needle or inserting a key into a lock. These tasks not only require precision and dexterity, but also an optimal camera perspective that captures and focuses on the relevant features of objects. We believe that a static camera may not always provide an optimal viewpoint, whereas a dynamic camera that adjusts its perspective in real-time to the task can offer more flexibility.

For example, consider the bimanual robotic task of inserting a key into a lock. If the lock or key is very small, fixed cameras positioned far from the scene may struggle to provide a clear view, making them difficult to locate and interact with. Additionally, depending on how a robot grasps the lock and key, the view of these objects from the fixed cameras may be occluded by the robot. When attempting to insert the key, the fixed camera may not be positioned at the right angle to see the keyhole. Eye-in-hand cameras attached to the robot's end-effectors can be ineffective since the camera viewpoints are dependent on the robot's task execution. In the example of inserting the key, the eye-in-hand cameras might not have a clear line of sight to the keyhole if the hole is obstructed by the key or the fingers.

To address problems of occlusion and poor perspectives, we propose using active vision (AV), adjusting one's viewpoint to find a more favorable perspective. This is inspired by how we handle manipulation tasks in our daily lives. For example, if a human were to insert a key into a lock, they might move their head and focus their gaze to one side (depending on the active hand) to more precisely manipulate the lock and key. How humans move their gaze independent of their arms, a well studied aspect of human visual feedback [10, 11, 12], is key to being able to manipulate everyday objects. As we move towards the ultimate goal of human level dexterity for robots, we hypothesize that similar adaptive visual feedback requirements should apply. Many research studies have integrated AV into applications like object tracking or scene reconstruction, demonstrating numerous benefits, including avoiding occlusions, overcoming limited fields of view, and focusing on key points of interest [13, 14]. However, in the context of robot learning and manipulation, research on AV remains underexplored.

In our work, we apply and evaluate AV-enhanced imitation learning for dexterous manipulation. Taking advantage of how humans can determine effective viewpoints for completing tasks, we develop a robot system where finding the best camera perspective is directly learned from human demonstrations. We build on the existing ALOHA 2 system [15], which has two robot arms for

bimanual manipulation, and introduce AV-ALOHA, which incorporates an additional 7 DoF arm (AV arm) carrying a stereo camera, dedicated solely for AV. During demonstration, the AV arm is controlled by the user's head and body movements to dynamically adjust the camera perspective. The user wears a VR headset that streams a live feed from the camera attached to the AV arm, offering an immersive, first-person active sensing experience. During training and data collection, the human operator seamlessly adjusts the camera's position using the AV arm by naturally moving their body, head, and neck. This allows them to simultaneously execute the task while attempting to capture an ideal perspective, independent of the manipulator arms. This setup allows for flexible camera movement, mimicking how humans can move their heads to find the best viewpoint. We also developed a simulation environment where users can collect data with just a VR headset and a computer, eliminating the need for physical robot hardware while maintaining the same immersive experience. In addition, in line with the principles of ALOHA, we keep the system open-source and cost-effective, using affordable components and robots. The extension only adds an estimated $6,600 to the overall cost.

With our teleoperation system, we collect data on a variety of simulation and real-world bimanual manipulation tasks and evaluate a state-of-the-art imitation learning policy, ACT [1], with and without the 7-DoF AV arm. The tasks we chose to test are relatively more challenging compared to those explored in previous publications and may require higher precision as well as be influenced by the selection of camera perspectives. We provide an extensive evaluation of AV in imitation learning and conduct ablation studies highlighting the impact of AV with different camera configurations.

Our contributions are as follows:

1. AV-ALOHA, an open-source, low-cost teleoperation system based on ALOHA 2 featuring an additional 7-DoF AV arm, providing a real-time and immersive VR teleoperation experience.
2. An open-source simulation environment for AV-ALOHA, featuring new bimanual manipulation tasks and publicly available datasets of human demonstrations for those tasks.
3. Extensive evaluation of active vision and imitation learning across various simulated and real-world tasks.
4. Ablation studies highlighting the impact of different camera configurations in combination with AV.

2 Related Work

2.1 Active Vision

Active vision (AV) was first defined in [16], where a framework was introduced to more efficiently solve tracking with an active observer. Since then, extensive research has focused on AV, particularly in the domain of object tracking [17, 18, 19, 20]. One key area of interest of AV is view planning, which seeks to determine the best sequence of views for a sensor [14]. Much of this work has been applied to object reconstruction [21, 22], scene reconstruction [23, 24, 25], object recognition [26], and pose estimation [27].

In manipulation, reinforcement learning (RL) policies have been developed for AV, modeling it as a partially observable Markov decision process (POMDP) to handle object manipulation in occluded environments [28, 29]. In this context, AV can be modeled as a POMDP, as the RL agent receives limited observations in the form of images from a camera, which may not fully represent the state. In addition to actions related to manipulating objects, the agent also has actions that adjust the camera viewpoint. By incorporating AV, the hope is that the agent can control its camera and adjust its observations to better infer the state.

There are also other learning-based approaches that utilize AV in manipulation. An energy-based method has been proposed to select the next best view, using a 7-DoF camera attached to an arm to

reduce energy and minimize surprise [30]. Additionally, data-driven AV approaches for grasping focus on selecting perspectives that optimize the grasping policy [31]. Some works also explore synthetic viewpoint augmentation to scale data for imitation learning, but these methods are not truly AV—they simply aim to increase data rather than find better views [32, 33]. Our approach differs in that we focus on human-guided AV. Unlike methods that are either too general to provide task-specific information, or too specialized for tasks like grasping, we propose a scalable approach. By learning from human demonstrations, the teleoperator naturally controls the camera view to find the best perspective.

2.2 Teleoperation Systems for Data Collection

Having a robust robot system for collecting human demonstrations is crucial. Recent works and systems have explored innovative approaches to gather robot data. Some systems utilize leader-follower configurations for bimanual control [1, 8, 34]. Others employ VR-based pose estimation or exoskeletons for cartesian space control [35, 36]. Additionally, some systems focus on simplifying data collection by creating devices that do not require a robot [3, 9, 37]. Instead of using parallel jaw grippers, many opt for multifingered hands controlled via hand pose estimation or motion capture gloves [38, 39, 40].

None of the systems mentioned incorporate AV with independent control of camera perspectives. While some systems, like Open-Television [41], use immersive first-person teleoperation with a VR headset and an AV two-axis gimbal, they maintain a relatively constrained camera movement, and AV is not their primary focus. Additionally, some industry systems, particularly those with humanoid robots, feature gimbal systems for head movement that only adjust camera direction [42, 43]. Our work, however, focuses on AV in controlling perspective independent of hand movement. Unlike these other systems, which have limited ranges of motion and degrees of freedom for AV, our setup uses a dedicated robotic arm, allowing for camera movement in 6-dimensional space much like a human's ability in adapting perspective. This enables exploration of diverse viewpoints, a complicating problem but at the same time opening new potentials in handling more complex robotic tasks.

2.3 Imitation Learning

Imitation learning, which involves learning from expert demonstrations, has proven to be an effective approach for robot control. Numerous general architectures have emerged [1, 2, 4, 5] and there also exists many efforts to scale up robot data and human demonstrations [7, 44] for generalist language-conditioned policies, [45, 46, 47, 48]. Despite notable advancements, robot learning continues to face significant challenges. Occlusions and the manipulation of small components remain difficult even for state-of-the-art methods [1]. The reliance on fixed or eye-in-hand cameras has restricted the ability of robots to effectively handle a range of manipulation tasks. Our work seeks to overcome these limitations by integrating AV into imitation learning to enable robots to tackle new and conventionally difficult tasks. Beyond performance improvements, we aim to deepen our understanding of the challenges posed by high-DoF AV systems, and demonstrate its potentials. We further hope to encourage the robotics community to explore these challenges and contribute to the development of next-generation robotic systems with human-like, minimalist and adaptive vision capabilities.

3 AV-ALOHA: Description of the Robot System

Our teleoperation, data collection, and autonomous system is illustrated in Figure 2. The system features three Interbotix ViperX-300 6-DoF [49] robotic arms: two equipped with grippers for manipulation and an AV arm fitted with a ZED mini stereo camera [50], whose movements are controlled by the operator's head movement via a Meta Quest 2 or 3 VR headset [51]. The two manipulation arms can be controlled either using VR controllers or the original ALOHA leader

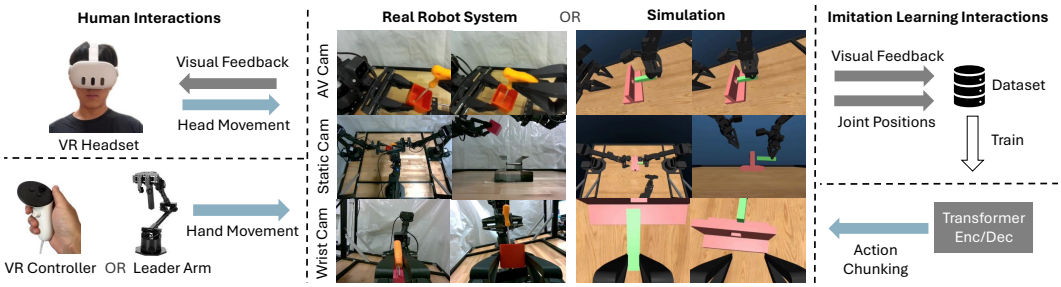

Figure 2: **Data collection and imitation learning pipeline with AV-ALOHA**: The AV-ALOHA system enables intuitive data collection using a VR headset for AV and either VR controllers or leader arms for manipulation (*left*). This helps capture full body and head movements to teleoperate both our real and simulation system that record video from six different cameras (*middle*) and provide training data for our AV imitation learning policies (*right*).

arms [52], which replicate joint positions. The camera on the AV arm streams two 720p RGB videos to the VR headset.

3.1 Hardware

The hardware configuration builds upon the ALOHA 2 [15] setup. We retain the two leader arms and the two follower arms, as well the original four Intel RealSense D405 cameras [53]. Two cameras are attached eye-in-hand to the follower arms while the other two are fixed to the top and bottom of the setup, providing high- and low-angle perspectives. We introduce a new arm, termed the AV arm, which is an Interbotix ViperX-300 arm equipped with a ZED mini stereo camera on its end effector. We further enhance this arm by converting it from 6-DoF to 7-DoF. The additional degree of freedom addresses the limited range of motion and frequent singularities encountered with the original 6-DoF configuration, significantly expanding the arm's ability to achieve various camera perspectives. This modification is straightforward, involving only the 3D printing of a small bracket and repurposing the existing gripper motor to provide a mechanism to easily pan the camera.

3.2 Simulation Environment

We also developed a simulation environment of the AV-ALOHA system using MuJoCo [54]. Building upon the ALOHA 2 model from MuJoCo Menagerie [55], we incorporated the AV arm, mirroring our real robot system. The data collection process in the simulation uses the same interface as the real robot system, as users can utilize a VR headset to stream stereo video and experience immersive teleoperation within the simulated environment. This simulation was created to offer a systematic and controlled setting for evaluating our AV and imitation learning policies.

3.3 Teleoperation with VR Headset

For the VR headset, we developed a Unity application that interfaces with the robot system via WebRTC [56]. The robot system streams two 720p, 30fps video feeds from the ZED mini's left and right cameras on the AV arm. These video streams are displayed independently to the operator's left and right eyes, enhancing immersion and providing a sense of depth and spatial awareness for the teleoperator.

We offer two teleoperation options, both involving the VR headset. The first option uses the VR headset exclusively for control. The headset transmits the tracked poses of the operator's head and hand controllers to the robot system, which then commands the arms accordingly. The grippers of the two arms are operated by pressing the trigger buttons on the hand controllers. The second option integrates the leader arms from the original ALOHA 2 system, allowing for control of the follower arms while the VR headset manages the AV arm. We provide these two interfaces for convenience. For simulation data collection, we chose the VR-only control option for its simplicity and lack of

additional hardware requirements. For real-world data collection, we opted for the leader arms due to reduced operator fatigue and better joint-wise control.

For the VR headset, we obtain the absolute poses of the user's head and hands and convert them to the robot's coordinate frame. The robot arms are initialized at a starting position, and if the first teleoperation option is used, the operator is provided with a visual AR guide for hand placement. Once control begins, all movements are relative to this initial pose. For both teleoperation options, the AV arm receives a target pose from the VR headset. Differential Inverse Kinematics (IK) with Damped Least Squares [57] is used to map this pose to the arm's joint angles. For controlling the manipulation arms with VR hand controllers, we use a Differential IK method with a custom cost function due to these arms frequently approaching joint singularities. This approach evaluates different joint deltas to find those that minimize the cost function. Our cost function penalizes deviations of joints from their center to avoid joint limits and reduces overall joint displacement to prevent overly sharp movements.

With our system, attaching the camera to a 7-DoF arm allows for an extended reach and enhanced range of motion. This setup enables us to precisely map the locations of the robot's two end-effectors and the AV arm to correspond exactly with the operator's hand and head positions. This alignment creates a more immersive experience, as the operator and robot directly mirror each other's movements. We believe this provides a more intuitive control and teleoperation experience, making it easier for users to learn and adapt to the system and hence, generate more natural and effective demonstration data for robot learning.

4 Experiments

To evaluate the effectiveness of AV for imitation learning, we adopt a popular imitation learning framework, Action Chunking with Transformers (ACT) [1]. We train and evaluate ACT using the library, LeRobot [58], which provides a state of the art implementation of ACT. We use the default implementation of ACT from LeRobot which uses a pretrained ResNet18 [59] visual backbone. For action chunking, we use a chunk size of 50 for both the simulation and real-world since the real-world data is collected at 50 Hz. Although the robot is teleoperated with cartesian control inputs, we record and train on joint position observations and actions. For training, we incorporate a learning rate of 2.5e-5 with a batch size of 16. We train for a total of 15625 steps. All other model parameters match the default configuration provided by LeRobot. During training, instead of relying on validation loss, we save multiple checkpoints and directly evaluate the policy on the target environment [60].

We evaluate on five simulation tasks and one real-world task. These tasks are a mix of bimanual tasks with varying levels of difficulty. For each task we experiment with different combinations of cameras for the model. AV-ALOHA has six cameras where two are fixed, two are attached eye-in-hand to the wrists of the arm, and two are from the stereo camera attached to the AV arm. We refer to the fixed cameras as **Static**, eye-in-hand cameras attached to the wrist as **Wrist**, and AV camera as **AV**. We experiment on all 7 different possible combinations of these cameras and evaluate success rates on the tasks.

For each task, we collected 50 episodes of human demonstrations with all three arms while recording from all cameras. Then we selectively train with the specific camera configuration. We decided not to collect separate data for each different camera configuration to keep the trajectories of the data consistent for training between different configurations. However, for configurations that don't require an AV camera, we acknowledge that the AV arm could potentially show up in the frame of the other cameras while also requiring additional control inputs for the policy. Thus, in the real world experiment, even when the policy doesn't use the AV camera we still have it control the AV arm to keep the data consistent. However for simulation experiments, we can render and record the same trajectories twice to both include and not include the AV arm. Therefore, in simulation, we can train and evaluate with and without the AV arm.

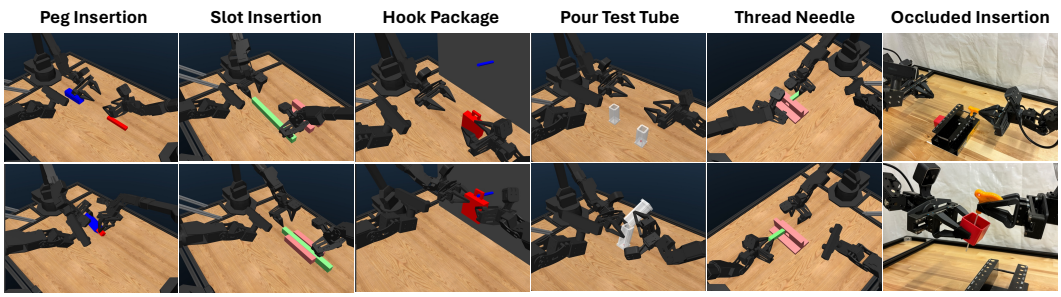

Figure 3: We experimented with five simulation tasks and one real-world task, each with varying levels of difficulty. Some tasks encourage the robot to actively seek optimal perspectives for successful execution.

4.1 Tasks

All six tasks, depicted in Figure 3, are designed to conduct bimanual manipulation. Three of these tasks (Group 1)—**Peg Insertion**, **Slot Insertion**, and **Hook Package**—can be completed without the need for AV, as the standard ALOHA camera setup (including static and wrist cameras) is sufficient for task execution. In contrast, the remaining three tasks (Group 2)—**Pour Test Tube**, **Thread Needle**, and **Occluded Insertion**—are designed to potentially benefit from improved camera perspectives provided by AV. By evaluating both scenarios, we gain insight into the advantages of AV in the latter group, while, through the former, we also identify any potential drawbacks, such as increased complexity from operating in a larger action space, or processing additional cameras inputs.

Peg Insertion is a simulation task adapted from the original ALOHA paper [1] to our new simulation where the right arm needs to grasp a peg/stick and the left arm needs to grab a socket. The two arms then coordinate to insert the peg into the socket. **Slot Insertion** is a simulation task adapted from [4] where the two arms need to grasp a long stick from both ends and insert it into a slot. **Hook Package** is a new simulation task in which both arms work together to grasp a package or box. The package has a tab with a hole, and the objective is to hang the package on a hook attached to a wall. We categorize **Peg Insertion** and **Slot Insertion** into Group 1, as these tasks have previously been demonstrated autonomously using a single static camera, as shown in [1, 4]. **Hook Package** is also placed in Group 1, as it is designed so that the package and hook remain clearly visible to either the static or wrist cameras.

Pour Test Tube is a new simulation task involving two slim tubes, where one tube contains a small marble. The two arms need to grasp the tubes and the right arm pours the marble from one tube into the other tube. We believe this task may require AV to look closely at the ball while pouring for more precision, similar to how a human might accomplish this task. **Thread Needle** is a new simulation task where the right arm grasps a needle and threads it through a hole on an object for the left arm to grab and pull it out on the opposite side of the hole. The object with the hole is placed so that its hole has limited visibility from the static cameras in the setup. **Occluded Insertion** is a real-world task where the right arm grabs a long allen key from a tray and the left arm grabs a small container with a hole at the bottom of the container. The right arm then needs to insert the end of the allen key into the hole of the container. However, since the hole is located at the bottom of the container, it may be occluded by the sides of the container, making the hole difficult to see without a proper camera perspective adjustment. These three tasks are categorized into Group 2 because they involve occlusions or require a focus on small details that could be better addressed with AV.

5 Results and Discussion

For simulation tasks, we evaluate each camera configuration using 12 different policy checkpoints from training, rolling out each checkpoint 50 times, and report the results for the best-performing

	Group 1						Group 2					
	Peg Insertion		**Slot Insertion**		**Hook Package**		**Pour Test Tube**		**Thread Needle**		**Occluded Insertion**	
	Grasp	Insert	Grasp	Insert	Grasp	Hook	Grasp	Pour	Grasp	Thread	Grasp	Insert
AV	74	42	88	50	100	22	66	**14**	98	**52**	60	20
AV + Static	84	46	100	62	100	34	50	10	98	26	20	0
AV + Wrist	82	34	96	44	100	22	70	**14**	92	**52**	95	**30**
AV + Static + Wrist	78	36	100	36	100	24	36	8	90	40	40	5
Static	84	**48**	98	66	100	**44**	44	8	88	30	85	20
Static + Wrist	88	40	100	**78**	100	30	46	6	38	22	100	15
Wrist	84	42	98	44	92	8	44	10	94	44	60	15

Table 1: Success rates (%) of the ACT policy across different tasks and camera configurations in simulation and real-world settings. Each task consists of two steps: "Grasp" (pickup objects in task) and a task-specific action completing the task, which indicate partial and full success respectively. Group 1 includes tasks designed to not necessarily require AV, while Group 2 includes tasks with occlusions or requiring high precision that could benefit from AV.

checkpoint. For the real-world task, we roll out the final checkpoint 20 times and report the success rate for each camera configuration. The results are presented in Table 1.

For group 1 tasks, non-AV setups achieved higher success rate on two tasks, (*slot insertion* and *hook package*), and for the *peg insertion* task, results were comparable between the two setups. This indicates that for those cases where AV is not necessarily advantageous, inclusion of additional camera feeds can deteriorate performance. This is further indicated when comparing the results of camera combinations within the non-AV scenarios. For example for *hook package*, while the highest performance is obtained when using static cameras alone, inclusion of the wrist cameras adversely affects the success rate.

For group 2 tasks, we found that setups with AV showed improvements over non-AV setups. For *thread needle* and *pour test tube*, camera configurations with AV performed exceptionally well, with the AV and AV + wrist configurations achieving the two highest success rates on these tasks. For *thread needle*, the hole is not easily visible from the static cameras and the AV camera is able to get a good perspective of the hole, which is crucial for inserting the needle through the hole. When comparing the perspectives of the wrist and AV cameras, although wrist cameras are able to provide a reasonable view of each side of the hole, the AV camera can provide a more holistic view of all the components involved. For *pour test tube* due to the slim design of the tubes and small marble size, the AV camera can better focus and zoom in on the openings of the tubes, and further provide a clearer view of the marble for better precision in the task.

For our real-world task, *occluded insertion*, the AV + wrist combination performed exceptionally well compared to other configurations. During inference with AV + wrist, we observed that the arm with the peg appeared to use visual feedback to align the peg and the hole, whereas in configurations without AV, the arm with the peg would forcefully press the peg against the container. From this behavior, we infer that AV enhances precision in tasks where visual feedback of small and intricate details is crucial.

Although setups with AV may not always achieve the best performance, we found that they consistently perform reasonably well across all tasks. We observed that AV alone without any fixed cameras achieved the top success rate on *pour test tube* and *thread needle* and performed relatively well for the other tasks. We can infer from this result that AV might be sufficient enough for decent performance across multiple tasks, without the need to install multiple fixed cameras around the scene.

Results indicate that using static cameras only performed best on two of the group 1 tasks, i.e. *peg insertion*, and *hook package*. Investigation of the camera feeds indicated that in these tasks, static camera perspective provided all the necessary visual information to complete the task. Additionally, since these cameras do not move, they provide a more stable and predictable vision input, whereas AV and wrist camera images change significantly following the control inputs, resulting in a more complex visual feedback/control system. We also hypothesize that fixed cameras benefit from a

"fixed coordinate system" where objects in a particular location are always located in the same corresponding pixel positions, making it easier for the model to interpret their locations. In contrast, the moving cameras, would introduce additional complexity in tracking and interpreting object locations on the scene. In such cases, where static cameras suffice to execute the task, we noticed that adding more moving cameras can deteriorate the results.

We further observed that in 4 out of 6 tasks, AV + static combination, outperforms static + wrist. These two configurations are similar in terms of data complexity and network architecture, but in AV + static scenarios the perspective control is decoupled from the object manipulation, unlike static + wrist. It is noted that AV + static setup further complicates the control requirements and hence, may gain additional benefit from more complex control architectures.

Another observation is that when a camera on its own achieves poor performance, adding that camera to the setup will significantly drop the performance. This phenomenon is observed in tasks *pour test tube*, *hook package*, *occluded insertion* and *thread needle*. For example in *thread needle*, adding static camera to AV dropped the success rate from 52% to 26%.

Another interesting result was that using all the cameras simultaneously did not perform well across the tasks and never ranked in the top three for any task. One explanation from the observations is that adding more cameras can actually hurt performance if the additional cameras do not provide significant new information. This result can be further attributed to the significantly larger action space and more complex, decoupled nature of vision control in AV, which may necessitate more complex control architectures, more training data, or additional training.

We further observed that finding a single consistent optimal camera perspective during training data generation led to the best performance, as it shrinks the solution space making it easier for the imitation model to learn.

6 Conclusion

In this work, we introduced a novel robotic setup featuring a 7-DoF AV arm, which extends the ALOHA 2 system. Through extensive experiments, we demonstrated that AV can significantly improve imitation learning, particularly in tasks that can benefit from proper selection of the camera perspective. Our results suggest that AV may be sufficient to provide the necessary visual feedback for successful task execution across a broad range of tasks, thereby improving the generalizability of robotic platforms and potentially reducing the need for additional, task-specific camera setups. This approach is inspired by how humans dynamically adjust their perspectives using head, neck, and waist movements to optimize their view during manipulation tasks.

To explore the utility of AV, we conducted an in-depth analysis of its role alongside static and conventional moving (eye-in-hand) camera setups. Our findings revealed that adding more cameras does not necessarily enhance performance, and in some cases, may even complicate the system. While AV shows significant potential, it also introduces complexities that warrant further research, particularly in the development of control architectures capable of managing the decoupled nature of visual feedback and control, the expanded action space, and the method's susceptibility to distribution shift. Additionally, our results raise important questions regarding the data and training requirements for such systems, highlighting the need for continued investigation into how AV can be integrated into robotic platforms.

Moving forward, we aim to contribute to the development of more generalizable, human-like robotic platforms where task-specific visual data is delivered in a targeted, controlled manner, with redundant information filtered at the sensing level rather than at the computational level. The results presented in this paper, along with the introduction of the open-source AV-ALOHA hardware and software, represent important first steps toward achieving this goal. We hope that this work inspires further research and development in this direction, ultimately leading to more efficient and adaptable robotic systems.

Acknowledgments

We thank Mohnish Gopi for assisting in the development of the VR teleoperation app and Soumyajit Ganguly for helping design 3D prints for the tasks.

References

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *RSS*, 2023.

[2] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

[3] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: Inthe-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[4] A. Lee, I. Chuang, L.-Y. Chen, and I. Soltani. Interact: Inter-dependency aware action chunking with hierarchical attention transformers for bimanual manipulation. *arXiv preprint arXiv:2409.07914*, 2024.

[5] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024.

[6] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[7] O.-X. E. Collaboration, A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[8] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[9] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[10] J. M. Findlay and I. D. Gilchrist. *Active vision: The psychology of looking and seeing*. Number 37. Oxford University Press, 2003.

[11] G. Maiello, M. Schepko, L. K. Klein, V. C. Paulun, and R. W. Fleming. Humans can visually judge grasp quality and refine their judgments through visual and haptic feedback. *Frontiers in Neuroscience*, 14:591898, 2021.

[12] R. Goodman and L. Tremblay. Using proprioception to control ongoing actions: dominance of vision or altered proprioceptive weighing? *Experimental Brain Research*, 236:1897–1910, 04 2018. doi:https://doi.org/10.1007/s00221-018-5258-7.

[13] S. Chen, Y. Li, and N. M. Kwok. Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research*, 30(11):1343–1377, 2011.

[14] R. Zeng, Y. Wen, W. Zhao, and Y.-J. Liu. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 6:225–245, 2020.

[15] ALOHA 2 Team. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation, 2024. URL `https://aloha-2.github.io/`.

[16] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International journal of computer vision*, 1:333–356, 1988.

[17] E. Rivlin and H. Rotstein. Control of a camera for active vision: Foveal vision, smooth tracking and saccade. *International Journal of Computer Vision*, 39:81–96, 2000.

[18] Denzler, Zobel, and Niemann. Information theoretic focal length selection for real-time active 3d object tracking. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 400–407. IEEE, 2003.

[19] M. Jiang, R. Sogabe, K. Shimasaki, S. Hu, T. Senoo, and I. Ishii. 500-fps omnidirectional visual tracking using three-axis active vision system. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021.

[20] Y. Liu, P. Sun, and A. Namiki. Target tracking of moving and rotating object by high-speed monocular active vision. *IEEE Sensors Journal*, 20(12):6727–6744, 2020.

[21] M. Krainin, B. Curless, and D. Fox. Autonomous generation of complete 3d object models using next best view manipulation planning. In *2011 IEEE international conference on robotics and automation*, pages 5031–5037. IEEE, 2011.

[22] A. K. Burusa, E. J. van Henten, and G. Kootstra. Attention-driven next-best-view planning for efficient reconstruction of plants and targeted plant parts. *Biosystems Engineering*, 246: 248–262, 2024.

[23] A. J. Davison and D. W. Murray. Simultaneous localization and map-building using active vision. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):865–880, 2002.

[24] A. J. Davison, W. W. Mayol, and D. W. Murray. Real-time localization and mapping with wearable active vision. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pages 18–27. IEEE, 2003.

[25] S. Dong, K. Xu, Q. Zhou, A. Tagliasacchi, S.-Q. Xin, M. Nießner, and B. Chen. Multi-robot collaborative dense scene reconstruction. *ACM Transactions on Graphics*, 38:1–16, 07 2019. doi:10.1145/3306346.3322942.

[26] B. Browatzki, V. Tikhanoff, G. Metta, H. H. Bülthoff, and C. Wallraven. Active object recognition on a humanoid robot. In *2012 IEEE international conference on robotics and automation*, pages 2021–2028. IEEE, 2012.

[27] K. Wu, R. Ranasinghe, and G. Dissanayake. Active recognition and pose estimation of household objects in clutter. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4230–4237. IEEE, 2015.

[28] R. Cheng, A. Agarwal, and K. Fragkiadaki. Reinforcement learning of active vision for manipulating objects under occlusions. In *Conference on Robot Learning*, pages 422–431. PMLR, 2018.

[29] Y. Fujita, K. Uenishi, A. Ummadisingu, P. Nagarajan, S. Masuda, and M. Y. Castro. Distributed reinforcement learning of targeted grasping with active vision for mobile manipulators. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9712–9719. IEEE, 2020.

[30] T. Van de Maele, T. Verbelen, O. Çatal, C. De Boom, and B. Dhoedt. Active vision for robot manipulators using the free energy principle. *Frontiers in neurorobotics*, 15:642780, 2021.

[31] S. Natarajan, G. Brown, and B. Calli. Aiding grasp synthesis for novel objects using heuristic-based and data-driven active vision methods. *Frontiers in Robotics and AI*, 8:696587, 2021.

[32] L. Y. Chen, C. Xu, K. Dharmarajan, Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024.

[33] S. Tian, B. Wulfe, K. Sargent, K. Liu, S. Zakharov, V. Guizilini, and J. Wu. View-invariant policy learning via zero-shot novel view synthesis. *arXiv preprint arXiv:2409.03685*, 2024.

[34] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. *arXiv preprint arXiv:2309.13037*, 2023.

[35] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik. Learning visuotactile skills with two multifingered hands. *arXiv preprint arXiv:2404.16823*, 2024.

[36] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang. Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. *arXiv preprint arXiv:2408.11805*, 2024.

[37] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu. Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15031–15038. IEEE, 2024.

[38] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[39] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

[40] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024.

[41] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[42] Reachy by Pollen Robotics, an open source programmable humanoid robot, . URL https://www.pollen-robotics.com/.

[43] Sanctuary ai, . URL https://sanctuary.ai/.

[44] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[45] 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.

[46] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. 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.

[47] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[48] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[49] ViperX 300 S, . URL https://www.trossenrobotics.com/viperx-300.

[50] ZED Mini Stereo Camera | Stereolabs, . URL https://www.stereolabs.com/store/products/zed-mini.

[51] Meta Quest VR Headsets, Accessories & Equipment | Meta Quest, . URL https://www.meta.com/quest/.

[52] WidowX 250 S, . URL https://www.trossenrobotics.com/widowx-250.

[53] Depth Camera D405, . URL https://www.intelrealsense.com/depth-camera-d405/.

[54] 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*, pages 5026–5033. IEEE, 2012. doi:10.1109/IROS.2012.6386109.

[55] K. Zakka, Y. Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL http://github.com/google-deepmind/mujoco_menagerie.

[56] aiortc/aiortc, Sept. 2024. URL https://github.com/aiortc/aiortc. original-date: 2018-02-23T22:05:16Z.

[57] S. R. Buss. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. *IEEE Journal of Robotics and Automation*, 17(1-19):16, 2004.

[58] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, and T. Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

[59] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[60] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.