

# AVR: Active Vision-Driven Robotic Precision Manipulation with Viewpoint and Focal Length Optimization

Yushan Liu<sup>\*†1</sup>, Shilong Mu<sup>\*‡1</sup>, Xintao Chao<sup>1</sup>, Zizhen Li<sup>2</sup>, Yao Mu<sup>3</sup>, Tianxing Chen<sup>4</sup>, Shoujie Li<sup>1</sup>, Chuqiao Lyu<sup>†1</sup>, Xiao-Ping Zhang<sup>1</sup>, *Fellow, IEEE*, Wenbo Ding<sup>†1</sup>

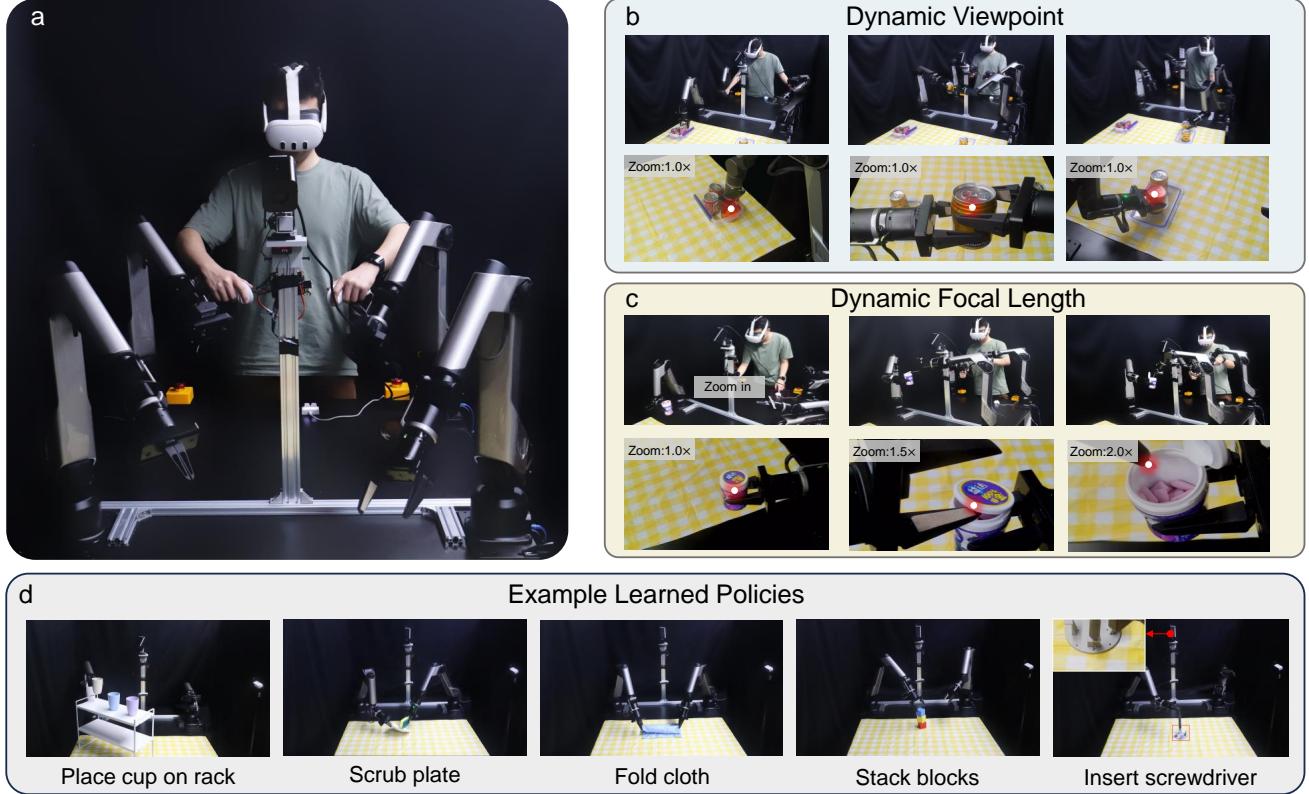


Fig. 1: Overview of the AVR system. (a) System setup with active vision, which can use a VR-controller or teleoperation solution to control robotic arms. (b-c) Demonstration of dynamic viewpoint and focal length adjustments during data collection for various tasks. The small keyboard is used to adjust the focal length when using ALOHA mode. (d) Example of policies learned for various robotic manipulation tasks, including placing cups, scrubbing plates, folding cloth, stacking blocks, and inserting screwdriver.

**Abstract**—Robotic manipulation within dynamic environments presents challenges to precise control and adaptability. Traditional fixed-view camera systems face challenges adapting to change viewpoints and scale variations, limiting perception and manipulation precision. To tackle these issues, we propose the Active Vision-driven Robotic (AVR) framework, a teleoperation hardware solution that supports dynamic viewpoint and dynamic focal length adjustments to continuously center targets and maintain optimal scale, accompanied by a corresponding algorithm that effectively enhances the success

rates of various operational tasks. Using the RoboTwin platform with a real-time image processing plugin, AVR framework improves task success rates by 5%-16% on five manipulation tasks. Physical deployment on a dual-arm system demonstrates in collaborative tasks and 36% precision in screwdriver insertion, outperforming baselines by over 25%. Experimental results confirm that AVR framework enhances environmental perception, manipulation repeatability ( $40\% \leq 1 \text{ cm}$  error), and robustness in complex scenarios, paving the way for future robotic precision manipulation methods in the pursuit of human-level robot dexterity and precision. Project page: <https://AVR-robot.github.io>.

<sup>\*</sup>These authors contributed equally to this work.

<sup>†</sup>Corresponding authors.

<sup>1</sup> Tsinghua University

<sup>2</sup> National University of Singapore

<sup>3</sup> Shanghai Jiao Tong University

<sup>4</sup> The University of Hong Kong

## I. INTRODUCTION

Robotic teleoperation has shown great potential in industrial assembly [1], minimally invasive surgery [2], and hazardous environments [3]. However, traditional systems

relying on fixed-view single-focal-length cameras are limited in complex scenarios because workspace changes alter all viewpoints simultaneously and hinder the operator to capture the optimal target view. [4], [5]. Moreover, the visual system's inability to adapt to scale and spatial pose shifts reduces precision and hand-eye coordination [6]. While imitation learning improves the efficiency of policy acquisition, its adaptability in dynamic environments remains limited [7]. Therefore, developing a visual feedback system with adjustable viewpoints and focal lengths is crucial for enhancing robotic teleoperation precision [8], [9].

To address this, we propose the **Active Vision-driven Robotic (AVR)** framework, which integrates an active vision system capable of dynamically adjusting both the camera's viewpoint and focal length. This ensures that the target remains centered and at an optimal scale throughout the operation. We developed a teleoperation robotic system featuring a dual-arm robotic manipulator based on the Galaxaea A1 [10] platform and an advanced visual perception module. A high-precision active vision unit consists an industrial camera with electronically adjustable zoom and a two-axis pan-tilt mechanism for precise viewpoint control.

Our system utilizes a Meta Quest 3 headset [11] as the master terminal, enabling a bidirectional link with a 1080p, 60 Hz video stream and real-time head pose feedback. An optimized control algorithm drives a pan-tilt mechanism ( $\pm 0.5^\circ$  accuracy) and supports two modes: button-mapped zoom via a VR controller and ALOHA-based [12] optical zoom ( $1\times$ – $7\times$ ). Validated on a RoboTwin-based simulation [13], our active vision techniques—using YOLOv8 [14] for ROI detection and a transformer-based SRGAN with dynamic adjustments [15]—improved task accuracy by 5%–16% with DP3 models [16]. In physical tests, an Action Chunking Transformer (ACT)-based system achieved up to 87% success, and a high-precision screwdriver insertion task improved by 25% (36% success) using dynamic  $5\times$  zoom and gimbal control. Additionally, the framework reduced the end-effector's positioning error ( $40\% \leq 1 \text{ cm}$ ), confirming its effectiveness in enhancing fine robotic operations.

Our contributions are as follows:

- We proposed the AVR framework, which integrates active vision optimization for robotic precision manipulation, enhancing the adaptability and efficiency of teleoperation systems.
- We conducted a series of systematic experiments on the RoboTwin-based simulation platform to verify the effectiveness of the AVR framework in enhancing task deployment accuracy.
- We developed a robotic system that integrates augmented reality teleoperation, a dual-arm manipulator, and an intelligent visual perception module, thereby demonstrating the practical applicability of the AVR framework in real-world scenarios.

## II. RELATED WORK

### A. Precision Manipulation

Precision manipulation refers to a robot's ability to perform highly accurate, meticulous, and flexible tasks in complex environments [17], [18]. Research in this area focuses on high-precision control and adaptability to dynamic conditions. Early approaches relied on rigid mechanical design and model-driven control, using kinematic and dynamic modeling to achieve precise positioning and assembly in structured settings [19]. More recently, deep learning and reinforcement learning have improved robot adaptability in dynamic environments [20], [21], while advances in vision and tactile sensing have enabled micrometer-level precision in grasping, manipulation, and assembly [22]. Additionally, multirobot collaboration has allowed for more complex and coordinated precision tasks. Despite significant progress, challenges remain in multiscale operation integration, dynamic disturbance compensation, and low-latency interaction [23]. Future research should further improve real-time alignment of cross-modal information and enhance the robustness of robotic vision in unstructured environments to optimize precision manipulation capabilities.

### B. Imitation Learning

Imitation learning is a popular approach for training robotic systems to perform complex tasks [24], [25]. By observing human demonstrations, imitation learning algorithms can learn to mimic human behavior and achieve high task performance [21]. Imitation learning has been successfully applied in various domains, such as autonomous driving, robotic manipulation, and computer vision [26]–[28]. In teleoperation systems, imitation learning algorithms have been used to train robots to perform tasks based on human demonstrations [29]. These algorithms can learn from human demonstrations and adapt to new environments, thereby improving the efficiency and robustness of teleoperation systems [30].

### C. Active Vision Systems

Active vision systems have been widely applied in robotics, computer vision, and human-computer interaction [19]. These systems are characterized by their ability to actively control the viewpoint and focal length of the camera, thereby enhancing the system's adaptability and robustness [31]. Active vision systems have been used in various applications, such as object tracking, scene reconstruction, and visual servoing [6], [32], [33]. In the field of teleoperation, active vision systems have been employed to improve the operator's situational awareness and operational efficiency [34]. For example, active vision systems have been used to optimize the camera viewpoint and focal length in teleoperation tasks, thereby enhancing the operator's ability to perceive and interact with the environment [4].

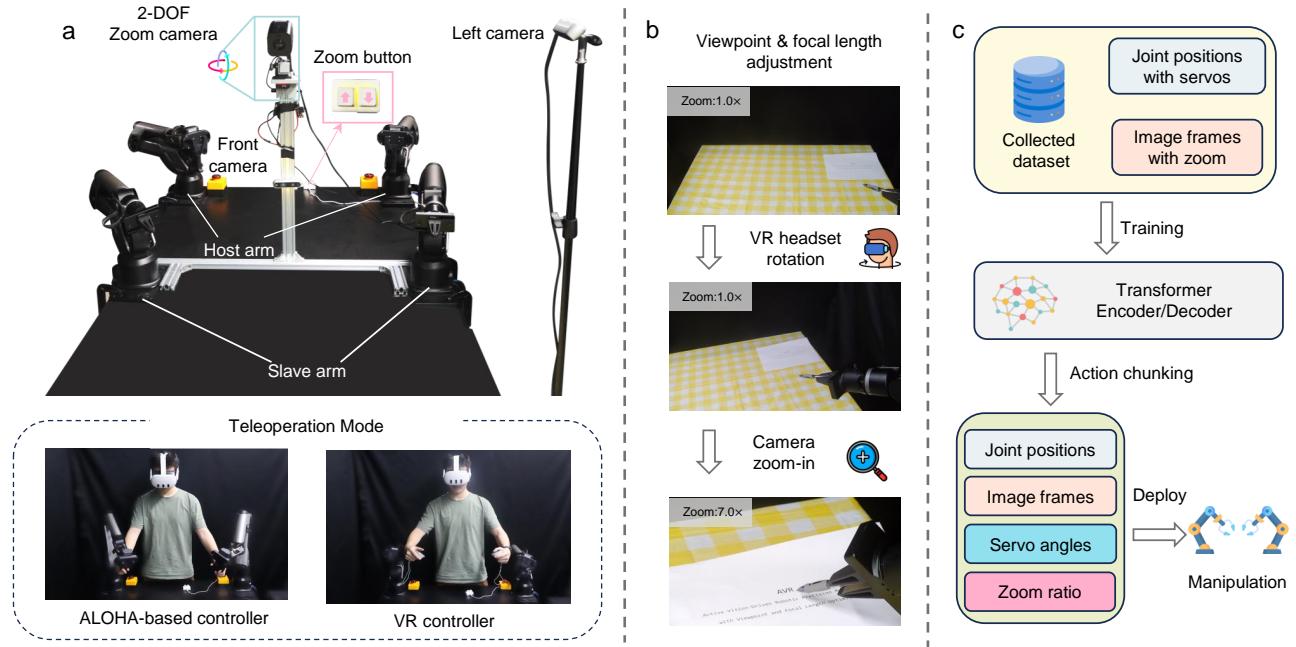


Fig. 2: AVR system architecture. (a) System have a 2-degree of freedom zoom camera, capable of covering the entire workspace, left camera and front cameras provide additional viewpoints. The system supports teleoperation via an ALOHA-based teaching pendant or a VR controller for intuitive manipulation.(b) The top camera enables 2D viewpoint adjustment and dynamic zoom. Users rotate the VR headset to center the target and adjust zoom via keyboard or controller. (c) Data-driven pipeline: collected joint positions, image frames, and zoom levels are processed by a transformer-based model for action chunking and deployed for manipulation.

### III. AVR SYSTEM ARCHITECTURE

#### A. Hardware System Overview

The system hardware architecture is shown in Fig. 2. We establish an interactive VR teleoperation platform [35], where the visual subsystem combines a distributed camera array with an active servo mechanism. Specifically, the central vision unit consists of a two-dimensional gimbal driven by dual servos (pitch range 0-60°, horizontal rotation range ±90°), supporting an industrial camera with an electric zoom function (4K resolution, 7× optical zoom, focal length range 4.8-48.2mm). The system achieves pose synchronization with the Meta Quest 3 headset through a quaternion-based pose mapping algorithm operating at up to 120 Hz. This algorithm processes the headset's IMU data to compute the required pan and tilt angles for the two-servo pan-tilt mechanism, ensuring that the target remains centered in a 1280×720 resolution image and enabling real-time observation and manipulation by the operator. To further extend the field of view, the system integrates two fixed Intel D435i cameras, creating a top-left-front three-camera layout, significantly enhancing the operational space coverage.

In terms of operational control, the system features two distinct interaction methods: (1) A precision mode is implemented using the industrial robotic arm teach pendant and Robot Operating System (ROS) to replicate the end-effector's pose at millimeter-level accuracy. In this mode, a two-

button keyboard allows for precise adjustments of the zoom ratio in increments of 0.05; (2) A VR controller mapping mode, which establishes a nonlinear mapping relationship between the controller's movements and the robotic arm's joint space, with dynamic zoom controls integrated into the controller buttons, optimizing the user experience. By leveraging VR devices for immersive control, the system refines the teleoperation experience with a more intuitive interface, improved efficiency, and a closer match to real-world interactions.

#### B. Data Collection and Learning Framework

As shown in Fig. 2 (b), the system framework integrates a comprehensive data acquisition workflow. Operators adjust the camera viewpoint using a head-mounted VR device, while precise zoom control is achieved via a keyboard or VR controller. We established an operational dataset that captures four key types of sensory data:

- Angular displacements of the 6-DOF mechanical arm joints and the gripper.
- Pitch and yaw angles from the dual-axis gimbal servo motors.
- Sequential image streams, including RGB-D data and optical metadata.
- Real-time dynamic zoom parameter matrices.

The training framework adopts an improved Action

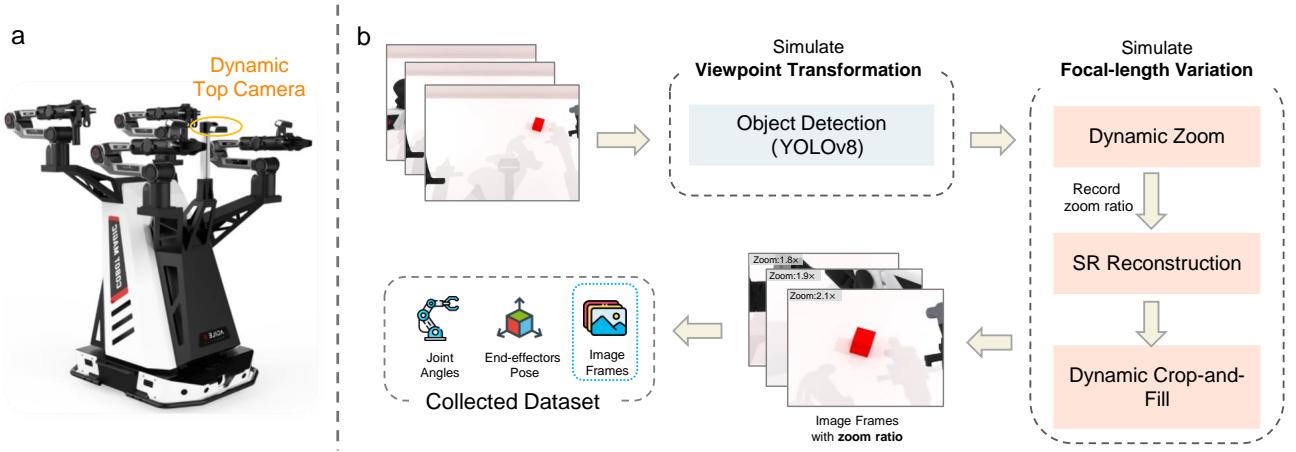


Fig. 3: RoboTwin-based viewpoint and focal-length variation simulation. (a) The RoboTwin-based simulation setup with the dynamic front camera. (b) Data extraction and processing flow: the collected dataset includes image frames, arm joint angles, and end-effector poses. Viewpoint transformation is simulated by object detection using YOLOv8, followed by dynamic zoom, super-resolution reconstruction and dynamic crop-and-fill operations ensure pixel consistency with the original dataset, enabling focal-length variation simulation. The processed data is then used for backend training.

Chunking Transformer (ACT) architecture based on **lertbot** [36]. After extensive training and deployment, the system demonstrates an intelligent control system capable of executing hierarchical motion planning.

### C. RoboTwin-based Simulation

To verify the theoretical feasibility of the AVR system, we developed a portable plugin within the RoboTwin simulation environment, enabling multi-scenario validation and deployment.

As shown in Fig. 3 (a), the data acquisition mechanism in RoboTwin employs a top-mounted camera with a global field of view, covering the entire operation space. To evaluate the impact of dynamic viewpoint adjustments and optical zoom on operation accuracy, we adopt the following strategy:

- Viewpoint transformation simulation: Real-time images from the top camera are processed to determine the target object based on the task requirements, with target detection and bounding box selection simulating dynamic adjustments of the camera's viewpoint.
- Optical zoom simulation: A super-resolution algorithm is employed to preserve the original image resolution after zoom adjustments, effectively emulating true optical zoom.

Our image processing pipeline comprises several stages that jointly enhance the target region for teleoperation. The workflow is detailed as follows:

YOLOv8-based object detection identifies the target and its bounding box. Let  $(x_{\min}, y_{\min})$  and  $(x_{\max}, y_{\max})$  be the bounding box coordinates, and define the target center as

$$x_c = \frac{x_{\min} + x_{\max}}{2}, \quad y_c = \frac{y_{\min} + y_{\max}}{2}.$$

We then apply an affine transformation to re-center the image by translating the target center to  $(W/2, H/2)$ :

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \quad T_x = \frac{W}{2} - x_c, \quad T_y = \frac{H}{2} - y_c.$$

This simulates a viewpoint adjustment so that the target is centered in the new image  $I_{vt}(x', y')$ .

**Optical Zoom Simulation:** A dynamic scaling factor  $s$  is computed to simulate optical zoom. The coordinates are scaled about the image center:

$$x'' = s(x' - \frac{W}{2}) + \frac{W}{2}, \quad y'' = s(y' - \frac{H}{2}) + \frac{H}{2}.$$

When  $s > 1$ , the ROI is enlarged (zoomed in). The resulting zoomed image,  $I_{zoom}$ , is then cropped appropriately.

**Super-Resolution Enhancement:** The cropped low-resolution target region  $\mathcal{P}_{LR} \in \mathbb{R}^{w \times h \times 3}$  is processed by a pre-trained Swin Transformer-based super-resolution network  $\mathcal{G}_\theta$ :

$$\mathcal{P}_{SR} = \mathcal{G}_\theta(\mathcal{P}_{LR}) = \mathcal{U}(\mathcal{B}(\mathcal{H}(\mathcal{P}_{LR}))).$$

We first extract features using a series of Transformer blocks with windowed self-attention, where the extracted features are computed as

$$\mathcal{H}(X) = \text{LN}(\text{MSA}(X) + X),$$

$$\text{MSA}(X) = \sum_{i=1}^h \text{Attention}(XW_Q^i, XW_K^i, XW_V^i)W_O^i.$$

with LN denoting layer normalization. Next, multi-level feature aggregation is performed through residual connections as follows:

$$\mathcal{B}(X) = X + \sum_{k=1}^K \text{SwinBlock}_k(X).$$

Finally, a pixel shuffle operation upsamples the features via

$$\mathcal{U}(X) = \text{PixelShuffle}(X * W_{\text{up}}),$$

where  $W_{\text{up}} \in \mathbb{R}^{C \times (r^2 C')}$  is the upsampling kernel and  $r$  represents the scaling factor.

**Output Format Verification and Iterative Refinement:** To maintain the original image format, a format verification matrix  $M_{\text{format}} \in \{0, 1\}^{H \times W}$  is defined:

$$M_{\text{format}}(i, j) = \begin{cases} \text{Bit Depth} = \text{BitDepth}(I_{\text{orig}}), \\ 1 & \text{if Color Space} \in \Gamma(I_{\text{orig}}), \\ & \text{Metadata} \equiv \text{Metadata}(I_{\text{orig}}) \\ 0 & \text{otherwise.} \end{cases}$$

An iterative correction is then performed:

$$I_{\text{final}}^{(k+1)} = I_{\text{final}}^{(k)} \odot M_{\text{format}} + I_{\text{orig}} \odot (1 - M_{\text{format}}).$$

All processed frames, together with supplementary data (additional camera views, robotic arm joint angles, gripper end-effector poses), are aggregated into a dataset for further training and deployment.

#### Algorithm 1: Image Frame Processing Pipeline

- 1: Given raw top-view image  $I_{\text{orig}}$ , max scaling factor  $S_{\text{max}}$ , safety coefficient  $\alpha$ .
- 2: Detect targets using YOLOv8, output bounding box  $(x, y, w, h)$ .
- 3: Compute scaling factor and record.
- 4: Feed into pre-trained Swin Transformer network  $\mathcal{G}_\theta$ .
- 5: Feature extraction, enhancement and image upsampling.
- 6: Validation matrix  $M_{\text{format}}$  and iterative correction.

## IV. EXPERIMENT

To comprehensively evaluate the capabilities of the AVR framework, we designed a series of experiments spanning both simulated and real-world environments. These experiments aim to assess the system’s performance across various robotic manipulation tasks, ranging from basic pick-and-place operations to high precision tasks requiring complex control strategies.

### A. Simulation Evaluation

On the RoboTwin-based simulation platform, we conducted a systematic experimental evaluation across five different tasks. Specifically, we collected 50 expert demonstration data for each task and trained the model based on DP3. Subsequently, we deployed in a simulated environment and assessed its task success rate, comparing the results with those of a baseline model. As shown in Table I, our approach achieved a success rate improvement ranging from 5% to 16% across all five tasks. These results further validate the effectiveness of our method in various task scenarios, demonstrating that dynamically adjusting visual focus and camera viewpoints can enhance the model’s ability to perceive key task-relevant objects.

TABLE I: Comparison of Task Success Rates between the Baseline and Our Method in RoboTwin-based Simulation

Task	Baseline	Ours
Block Hammer Beat	78%	89%
Block Handover	94%	99%
Blocks Stack(easy)	23%	39%
Container Place	54%	63%
Empty Cup Place	82%	95%

### B. Real Robot Performance

Based on our simulation results, We further deployed and evaluated our AVR system on a self-constructed physical experiment platform using two control schemes: a demonstration-based approach via an ALOHA teaching pendant and an interactive approach with VR controllers. The experimental tasks encompassed various bimanual pick-and-place operations—such as handover, folding clothes, and dish scrubbing—and some precision tasks designed to assess the impact of dynamic camera zooming, such as stacking three 5 cm-high blocks, grasping a 10 cm-long screwdriver to accurately insert it into a 1 cm-diameter hole (Fig. 4), grasping a  $2 \times 1$  cm piece of chewing gum from a 5-cm-diameter container. Table II summarizes the success rates obtained, comparing our AVR framework against baseline methods, thereby demonstrating the benefits of dynamically adjusting both the camera viewpoint and zoom for enhancing task performance.

We also analyzed how adjustments to the camera viewpoint and focal length impacted task success rates to varying extents. For some pick-and-place operations (e.g., handover or dish scrubbing), the top camera viewpoint undergoes substantial changes to keep the target in view, thereby improving success rates. However, because these tasks typically require only minor zoom adjustments (often less than  $2\times$ ) and do not involve high-precision operations, focal length variations have little effect on performance. By contrast, tasks such as folding cloth or placing cups exhibit minimal viewpoint and zoom changes, so introducing additional zoom parameters can introduce unnecessary complexity—sometimes reducing the success rate. Meanwhile, block stacking—requiring frequent zoom adjustments during the placement phase for precise alignment—exhibits a higher improvement in success rate when dynamic zoom is incorporated, compared to standard pick-and-place tasks. Finally, for precision tasks such as grasping chewing gum or inserting screwdriver, changing the viewpoint alone does not sufficiently clarify the target position; here, zoom adjustments become crucial in boosting success rates.

Overall, these results indicate that the AVR framework effectively improves task success rates across a variety of scenarios. Notably, for precision tasks, conventional imitation learning methods achieve very low success rates, whereas the AVR-enhanced model exceeds 25%. This underscores how our framework’s ability to capture and process finer details enables more complex and precise manipulations while improving stability and reliability.

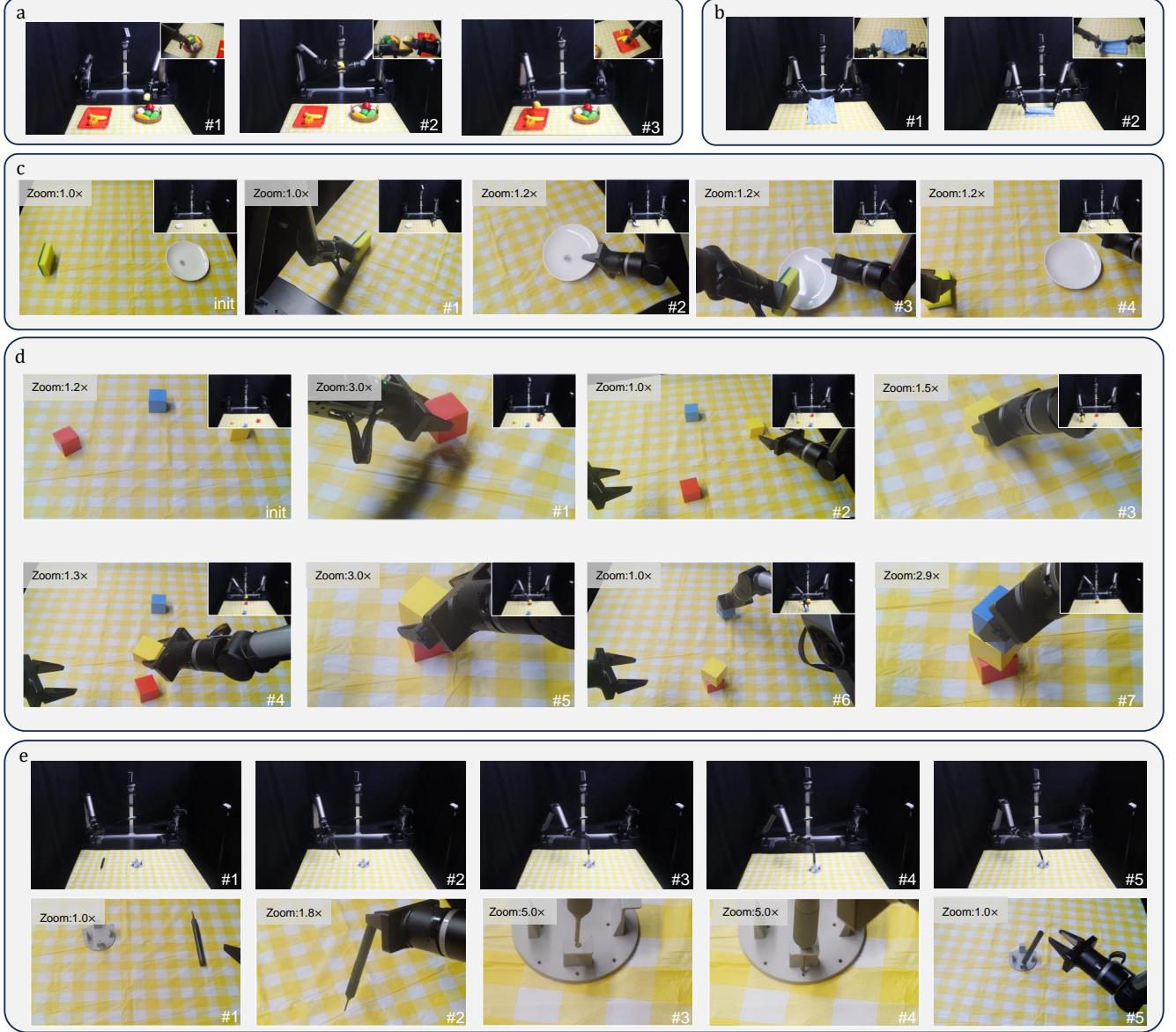


Fig. 4: Deployment of various manipulation tasks. (a) Pick-and-place of objects with varied shapes. (b) Folding fabric with coordinated bimanual manipulation. (c) Dish scrubbing with a controlled wiping motion. (d) Block stacking requiring precise alignment. (e) Inserting a screwdriver tip into a hole for assembly. We provide first-person views from the active vision camera at various stages of each task, capturing changes in viewpoint and real-time focal adjustments.

TABLE II: Success rates (%) for robotic manipulation tasks under three camera settings: baseline (static view), dynamic viewpoint only, and AVR (dynamic viewpoint with zoom).

Task	Data Collection		Success Rate (%)		
	Avg. time (s)	Succ./Trials	Baseline	W/O Dynamic Zoom	AVR (Ours)
Pick-place	30.5	40/40	75	87	<b>90</b>
Dish scrubbing	29.6	38/40	74	82	<b>89</b>
Fold cloth	35.6	38/40	71	<b>75</b>	67
Place cup on rack	22.1	40/40	82	<b>93</b>	92
Block stacking	52.5	27/30	35	57	<b>71</b>
Grasp chewing gum	46.3	28/30	24	27	<b>43</b>
Insert screwdriver	70.2	26/30	11	13	<b>36</b>

### C. Manipulation Accuracy

In the field of robotic precision manipulation, an important performance metric is repeatability accuracy, which refers to the system's ability to maintain minimal deviations when executing the same task repeatedly. Traditional imitation learning models often struggle with stability in repetitive tasks, primarily due to limited training data and inconsistencies in teleoperation demonstration quality. These factors make it difficult for the model to achieve high stability in precision operations.



Fig. 5: Repeatability accuracy experience for AVR. We design a "dart throwing" test. After 50 demonstrations and 15 times deployment, results (right) show that 40% of the attempts had an error below 1 cm.

To evaluate the ability of our AVR system to learn stable precision operations, we designed a "Dart Throwing Expert" experiment, as illustrated in Fig. 5. In this experiment, a 0.5 mm ballpoint pen was mounted on the end effector of the robot gripper, and a target board composed of concentric circles with 1 cm spacing was mounted on the table. During data collection, we recorded 50 demonstration trajectories, each aimed precisely at the target's center. To enhance visual perception during data collection, we dynamically adjusted the optical zoom of the top camera to approximately 6x, ensuring that the field of view primarily captured the inner three rings. The collected demonstrations were then used to train the model, which was subsequently deployed on the robot for testing.

The right figure illustrates the distribution of 15 deployment attempts. Each concentric circle corresponds to a specific error range; the color-coded rings and their numeric labels indicate the assigned scores, decreasing from the center outward. The numbers inside each ring represent how many times landed in that ring. Final results showed an average score of 9.3, with more than 40% of the trials achieved a deviation of less than 1 cm. Achieving such performance with a limited dataset highlights the effectiveness of our AVR framework in enhancing the learning capabilities of the model for precision tasks. It improves the system's understanding of fine details, leading to greater execution stability and overall robustness in complex robotic operations.

### V. CONCLUSION AND FUTURE WORK

In this work, we introduce the AVR framework, which leverages dynamic viewpoint and focal length adjustments in active vision to enhance precise manipulation. The system provides an intuitive teleoperation experience and a reliable data collection workflow, ensuring consistent dynamic viewpoint and zoom adjustments that contribute to stable operation and improved control during precision tasks. It demonstrates improved precision in kinds of manipulation tasks. Simulation and real-world experiments show that AVR improves task success rates by 5% - 16%, with more than 25% increases in precision for precision tasks, significantly outperforming conventional imitation learning methods. In particular, our "dart-throwing" experiment (average score: 9.3, with 40% of trials landing within 1 cm of the target) validates the system's enhanced precision in repeated end-effector positioning, highlighting its potential for high-precision robotic manipulation.

Future work will address several key areas. First, we aim to boost data collection efficiency by refining the interaction design and hardware configuration in ALOHA mode. Second, we plan to enhance viewpoint control by adopting higher-precision gimbal motors and further refining the VR-to-camera mapping algorithm. Finally, we will explore integrating additional robotic platforms—such as those equipped with wrist-mounted cameras and end-effector pose sensing—to further enhance perception and adaptability for complex operations.

### ACKNOWLEDGEMENTS

This work was supported by National Key R&D Program of China (No.2024YFB3816000), Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197), and Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005).

### REFERENCES

- [1] K. Jung, B. Chu, S. Park, and D. Hong, "An implementation of a teleoperation system for robotic beam assembly in construction," *International Journal of Precision engineering and manufacturing*, vol. 14, pp. 351–358, 2013.
- [2] A. J. Hung, J. Chen, A. Shah, and I. S. Gill, "Telementoring and telesurgery for minimally invasive procedures," *The Journal of urology*, vol. 199, no. 2, pp. 355–369, 2018.
- [3] C. Gentile, G. Lunghi, L. R. Buonocore, F. Cordella, M. Di Castro, A. Masi, and L. Zollo, "Manipulation tasks in hazardous environments using a teleoperated robot: A case study at cern," *Sensors*, vol. 23, no. 4, p. 1979, 2023.
- [4] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," 2024.
- [5] A. García, J. E. Solanes, A. Muñoz, L. Gracia, and J. Tornero, "Augmented reality-based interface for bimanual robot teleoperation," *Applied Sciences*, vol. 12, no. 9, p. 4379, 2022.
- [6] D. Rakita, B. Mutlu, and M. Gleicher, "An autonomous dynamic camera method for effective remote teleoperation," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 325–333, 2018.
- [7] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*, pp. 785–799, PMLR, 2023.

- [8] Y. Chen, L. Sun, M. Benallegue, R. Cisneros-Limón, R. P. Singh, K. Kaneko, A. Tanguy, G. Caron, K. Suzuki, A. Kheddar, *et al.*, “Enhanced visual feedback with decoupled viewpoint control in immersive humanoid robot teleoperation using slam,” in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pp. 306–313, IEEE, 2022.
- [9] I. Chuang, A. Lee, D. Gao, and I. Soltani, “Active vision might be all you need: Exploring active vision in bimanual robotic manipulation,” 2024.
- [10] Galaxea, “Galaxea a1.” [https://galaxea.ai/Introducing\\_Galaxea\\_Robot/product\\_info/A1/](https://galaxea.ai/Introducing_Galaxea_Robot/product_info/A1/), 2024. Accessed: Feb 1, 2025.
- [11] Meta Platforms, “Meta quest 3.” <https://www.meta.com/quest/quest-3/>, 2023. Accessed: Feb 1, 2025.
- [12] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware,” 2023.
- [13] Y. Mu, T. Chen, S. Peng, Z. Chen, Z. Gao, Y. Zou, L. Lin, Z. Xie, and P. Luo, “Robotwin: Dual-arm robot benchmark with generative digital twins (early version),” 2024.
- [14] R. Varghese and M. Sambath, “Yolov8: A novel object detection algorithm with enhanced performance and robustness,” in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6, IEEE, 2024.
- [15] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, and W. Xu, “Improved srgan for remote sensing image super-resolution across locations and sensors,” *Remote Sensing*, vol. 12, no. 8, p. 1263, 2020.
- [16] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, “3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations,” 2024.
- [17] M. R. Cutkosky, *Robotic grasping and fine manipulation*, vol. 6. Springer Science & Business Media, 2012.
- [18] B. Siciliano, O. Khatib, and T. Kröger, *Springer handbook of robotics*, vol. 200. Springer, 2008.
- [19] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, “Active vision,” *International journal of computer vision*, vol. 1, pp. 333–356, 1988.
- [20] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [21] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.
- [22] W. Yuan, S. Dong, and E. H. Adelson, “Gelsight: High-resolution robot tactile sensors for estimating geometry and force,” *Sensors*, vol. 17, no. 12, p. 2762, 2017.
- [23] A. Iqbal, “Stabilization of delayed teleoperation systems using time domain passivity control,” 2007.
- [24] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, “A survey of robot learning from demonstration,” *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.
- [25] J. Urain, A. Mandlekar, Y. Du, M. Shafiqullah, D. Xu, K. Fragkiadaki, G. Chalvatzaki, and J. Peters, “Deep generative models in robotics: A survey on learning from multimodal demonstrations,” *arXiv preprint arXiv:2408.04380*, 2024.
- [26] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, “End to end learning for self-driving cars,” *arXiv preprint arXiv:1604.07316*, 2016.
- [27] 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.
- [28] L. Pinto and A. Gupta, “Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours,” in *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 3406–3413, IEEE, 2016.
- [29] P. Abbeel and A. Y. Ng, “Apprenticeship learning via inverse reinforcement learning,” in *Proceedings of the twenty-first international conference on Machine learning*, p. 1, 2004.
- [30] S. Schaal, “Is imitation learning the route to humanoid robots?,” *Trends in cognitive sciences*, vol. 3, no. 6, pp. 233–242, 1999.
- [31] V. D. Cong and L. D. Hanh, “A review and performance comparison of visual servoing controls,” *International Journal of Intelligent Robotics and Applications*, vol. 7, no. 1, pp. 65–90, 2023.
- [32] 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*, pp. 400–407, IEEE, 2003.
- [33] Davison, “Real-time simultaneous localisation and mapping with a single camera,” in *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 1403–1410, IEEE, 2003.
- [34] D. Zhu, T. Gedeon, and K. Taylor, ““moving to the centre”: A gaze-driven remote camera control for teleoperation,” *Interacting with Computers*, vol. 23, no. 1, pp. 85–95, 2011.
- [35] S. P. Arunachalam, I. Güney, S. Chintala, and L. Pinto, “Holo-dex: Teaching dexterity with immersive mixed reality,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5962–5969, IEEE, 2023.
- [36] H. Face, “Ierobot.” <https://github.com/huggingface/ierobot>, 2023. Accessed: March 1, 2025.