# AVR: Active Vision-Driven Precise Robot Manipulation with Viewpoint and Focal Length Optimization

Yushan Liu[*1], Shilong Mu[*1], Xintao Chao[1], Zizhen Li[2], Yao Mu[3], Tianxing Chen[4], Shoujie Li[1], Chuqiao Lyu[†1], Xiao-Ping Zhang[1], *Fellow, IEEE*, Wenbo Ding[†1]

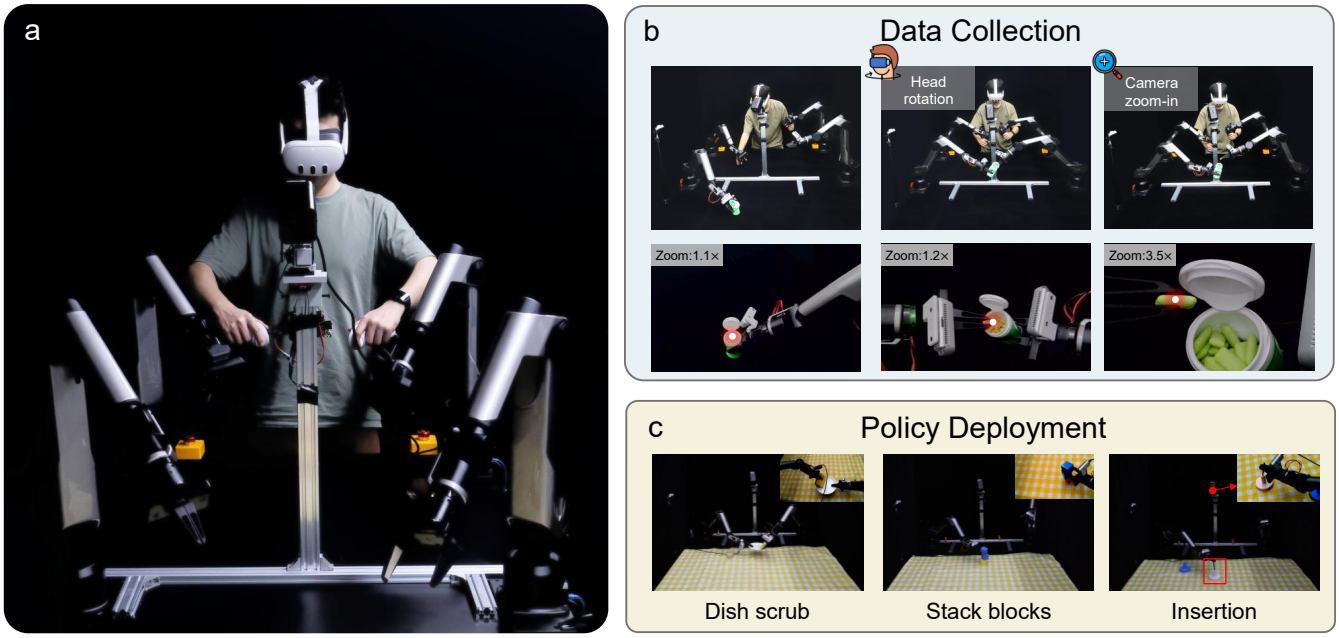


Fig. 1: **Overview of the AVR system**. (a) Hardware setup enabling bimanual control via VR controller or host–slave teleoperation. (b) Data collection with operator-controlled camera viewpoint (head pose) and focal length (zoom). (c) Example policy deployments: dish scrub, stack blocks, and insertion.

***Abstract*— Robotic manipulation in complex scenes demands precise perception of task-relevant details, yet fixed or suboptimal viewpoints often impair fine-grained perception and induce occlusions, constraining imitation-learned policies. We present AVR (Active Vision-driven Robotics), a bimanual teleoperation and learning framework that unifies head-tracked viewpoint control (HMD-to-2-DoF gimbal) with motorized optical zoom to keep targets centered at an appropriate scale during data collection and deployment. In simulation, an AVR plugin augments RoboTwin demonstrations by emulating active vision (ROI-conditioned viewpoint change, aspect-ratio-preserving crops with explicit zoom ratios, and super-resolution), yielding 5–17% gains in task success across diverse manipulations. On our real-world platform, AVR improves success on most tasks, with over 25% gains compared to the static-view baseline, and extended studies further demonstrate robustness under occlusion, clutter, and lighting disturbances, as well as generalization to unseen environments and objects. These results pave the way for future robotic precision manipulation methods in the pursuit of human-level dexterity and precision.**

*These authors contributed equally to this work.
†Corresponding authors.
[1] Tsinghua University
[2] National University of Singapore
[3] Shanghai Jiao Tong University
[4] The University of Hong Kong
Project page: `https://AVR-robot.github.io`.

## I. Introduction

Imitation learning (IL) has emerged as a powerful paradigm for enabling dexterous robotic behavior in complex systems. Unlike reinforcement learning, IL eliminates the need for precise dynamics modeling or the manual design of reward functions by directly learning end-to-end control policies from expert demonstrations. This approach has achieved remarkable progress across a range of robotic manipulation tasks [1]–[5]. However, as task complexity increases, for example in cluttered environments or in operations that require high-precision control, conventional IL frameworks continue to face significant performance bottlenecks.

Data collection for imitation learning in dexterous manipulation is hampered by teleoperation noise and perceptual limitations, which together yield suboptimal demonstrations and throttle policy progress. Third-person teleoperation typically operates with limited situational awareness and delayed

feedback, which often results in frequent errors or failed trials. This problem is especially evident in fine manipulation, and the inclusion of such demonstrations in training degrades the model's ability to capture essential action patterns. At the sensing level, conventional camera configurations, which combine fixed external views with wrist-mounted cameras, rarely provide the high-fidelity cues needed for precision. Fixed cameras offer global context but lack sufficient spatial resolution in cluttered or precision-sensitive scenes. Wrist cameras, on the other hand, provide local views that are often occluded and subject to erratic viewpoints as the manipulator moves. Consequently, even with state-of-the-art IL algorithms and large-scale, high-quality demonstrations, success rates on complex tasks often plateau, indicating a bottleneck rooted not merely in data quantity but in the incomplete capture and representation of critical interaction details during data collection, which ultimately limits policy generalization and robustness.

Existing work has make progress in VR-based teleoperation [6]–[8] and large-scale dataset building [9]–[12]. However, acquiring high-quality real-world demonstrations remains costly and time-consuming, and, as noted above, simply scaling data often yields diminishing returns, with success rates saturating on fine-grained manipulation tasks.

In contrast, human operators naturally leverage attention mechanisms to dynamically filter and focus perception on task-relevant regions and features. This adaptive allocation of sensory resources enables humans to consistently perceive and react to fine details, even under high cognitive load and visual clutter. Inspired by this, we posit that equipping robots with a similar capability to **actively focus on and magnify critical interaction details** during demonstration collection can effectively mitigate the aforementioned limitations.

To this end, we propose **Active Vision-Driven Robotics (AVR)**, a bimanual manipulation system that integrates an active vision module capable of dynamically adjusting both camera viewpoint and optical zoom. The framework consists of the following components:

- **Autonomous Optical Zoom Camera**: A controllable zoom camera to enable real-time magnification of task-relevant regions during demonstration. The recorded zoom information and video are jointly used as inputs for policy learning, providing an additional modality focused on fine-detail perception.
- **Egocentric VR with Head-Tracked View Control**: Stream the optical zoom camera to an VR head motion device(HMD) and map the operator's head pose to a 2-DoF gimbal for real-time viewpoint adjustment, keeping the target centered and yielding higher-fidelity, task-relevant demonstrations.

We evaluate our AVR framework across both simulation and real-world bimanual platform. Results demonstrate that our approach consistently improves task success rates across diverse scenarios, especially under cluttered or precision-demanding conditions. Notably, AVR achieves up to 30% improvement in success rate compared to state-of-the-art baselines under limited data conditions, validating its effectiveness in enhancing demonstration quality and policy performance through active visual attention.

## II. Related Work

### A. Active Vision for Robot Policy Learning

Active vision have been widely applied in robotics, especially for robot policy learning [14], [15]. Recent work mostly focus on "attention guidance" – either controlling camera viewpoints to reduce occlusion and field-of-view limitations, or jointly learning a sensory (viewpoint) policy with the motor policy in partially observable settings – thereby steering perception toward task-relevant regions and mitigating dilution in global views. [7], [16] In parallel, encoder-centric approaches (e.g., voxelized or two-stream encoders) emphasize spatially selective processing to highlight manipulation-relevant areas without explicitly moving the camera [17], [18]. However, these strategies often underrepresent fine, manipulation-critical cues, a gap made evident when fixed or suboptimal viewpoints induce occlusions and miss small details. In this work, we extend additional detail observation that dynamically adjusts viewpoint and zoom to capture critical details, enabling more precise and robust manipulation.

### B. Learning from Details

Learning from details is a recurring strategy for boosting perception and decision quality across domains. In fine-grained recognition, models mine subtle cues via region-level attention or localized feature interactions [19], [20]; In medical imaging, zoom-in pipelines explicitly crop high resolution patches around suspicious findings to capture lesion-level evidence [21]; For small-object detection, surveys and task-driven super-resolution demonstrate that recovering fine structure materially improves recognition in low resolution or cluttered scenes [22].

Robotics has also benefited from detail-centric perception [23]. Building on these insights, we go beyond implicit attention by introducing an explicit detail-observation modality. In practice, the system dynamically adjusts viewpoint and zoom to deliver high-resolution local evidence to the policy, thereby improving precision and robustness in fine manipulation.

## III. AVR System Architecture

### A. Hardware System

Fig. 2 shows the hardware architecture of our AVR system. The manipulation platform consists of two 6-DoF Galaxea robotic arms, each equipped with a parallel-jaw gripper [24]. Three Intel D435i depth cameras are positioned to capture side and wrist view, ensuring complete observation of the manipulation workspace. An active vision module is mounted on top of the platform, comprising a motorized-zoom industrial camera and a 2-DoF gimbal. It captures real-time human head motion with live visual feedback during teleoperation, which facilitates fine-grained perception and the recording task details.

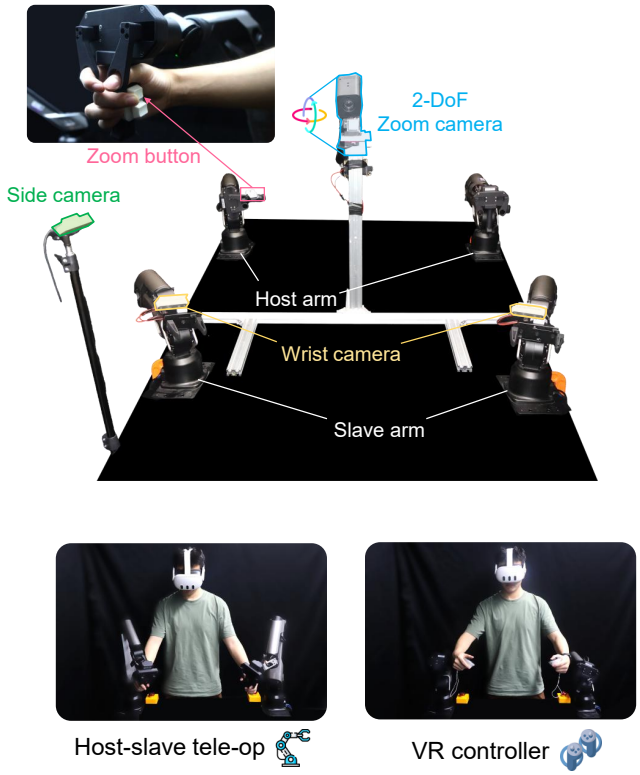


Fig. 2: AVR hardware platform. A 2-DoF zoom camera provides full-workspace coverage and adjustable focal length, complemented by a side camera and wrist-mounted cameras. System supports two control modalities: host–slave teleoperation and VR controller, each enabling real-time zoom control of the active camera (host–slave: dedicated zoom button; VR: controller button mapping).

### B. Teleoperation for Data Collection

To collect high-quality real-world manipulation data, we build a teleoperation framework with two control modes: (i) host–slave joint-angle mapping and (ii) 6-DoF end-effector control via VR controllers. This dual setup accommodates diverse tasks and operator preferences. For perception, we provide a VR-projected egocentric view with head-tracked active viewpoint control, closing a low-latency observation–action loop. We also add active zoom: a button–zoom mapping (keyboard button or VR controller) enables real-time focal adjustments during bimanual operation; at deployment, the policy autonomously modulates zoom to retain detail-rich observations.

***Pose-to-Gimbal Mapping***: Prior studies have demonstrated that 2-DoF gimbals can effectively approximate human head motion while maintaining adequate coverage of the operational workspace [25]. We leverage head motion estimates from HMD to capture head orientation, employing a One-Euro filter [26] for adaptive noise mitigation, and ultimately outputting gimbal rotation angles:

$$\boldsymbol{\theta}_t = \Pi(\,\mathbf{R}\,(\,\mathscr{F}_{\text{1-Euro}}(\tilde{\boldsymbol{q}}_t))) \tag{1}$$

where $\tilde{\boldsymbol{q}}_t \in \mathbb{S}^3$ is the raw unit quaternion at time $t$, $\mathscr{F}_{\text{1-Euro}}$ is the One-Euro filter defined on $\mathrm{SO}(3)$ with dynamically adjusted parameters:

$$f_t = f_{\min} + \beta\,\|\dot{\omega}_t\|\,, \quad \alpha_t = \frac{2\pi f_t \Delta t}{1+2\pi f_t \Delta t}, \tag{2}$$

where $\dot{\omega}_t$ is a low-pass estimate of angular velocity and $\Delta t$ is the sampling interval. $\mathbf{R}(\cdot)$ maps quaternion to rotation matrix. $\Pi(\cdot)$ extracts yaw/pitch from rotation matrix: letting $\boldsymbol{k} = \mathbf{R}_t^{\text{gimbal}}\mathbf{e}_z$, with $\boldsymbol{k} = [k_x, k_y, k_z]^\top$, $\mathbf{e}_z = [0,0,1]^\top$:

$$\begin{bmatrix}\theta_t^{\text{yaw}} \\ \theta_t^{\text{pitch}}\end{bmatrix} = \Pi(\mathbf{R}_t^{\text{gimbal}}) = \begin{bmatrix}\arctan(k_x, k_z) \\ \arctan\big(-k_y, \sqrt{k_x^2+k_z^2}\big)\end{bmatrix} \tag{3}$$

Compared with direct pose-to-gimbal mapping, the filter-based pipeline suppresses jitter more effectively at rest and delivers smoother tracking during rapid head motions, resulting a more stable view for teleoperation.

***Camera-to-HMD Projection***: To ensure a clear operational viewpoint, we adopt spherical rendering with an average end-to-end latency of approximately 80 ms. The live 60 fps camera stream is projected onto a 110° curved surface at a 1 m radius around the HMD viewpoint. By preserving vestibular–visual consistency, this spherical rendering can effectively mitigate motion sickness [27]. Combined with mentioned gimbal control, it provides a high-fidelity immersive perspective for precise teleoperation tasks.

***Button-to-Zoom Camera Control***. To better focus on and record fine interaction details during data collection, we equip both teleoperation modes with a button-to-zoom mapping for real-time control of the active vision camera. In the host–slave teleoperation, the operator uses a keypad on the teach pendant to zoom in/out; in the VR mode, controller buttons are mapped to camera zoom, enabling simultaneous bimanual operation and real-time adjustment of field of view and focal length. At deployment, the camera's focal length is autonomously adjusted according to the policy's learned zoom behavior, ensuring detail observation during from collection to execution.

### C. Learning Policy

We designed a policy network based on Diffusion Policy [4]. By leveraging external and active vision observations, along with proprioceptive state, the network predicts control actions for the system.

At each time step $t$, the policy receives the current RGB image observations $\mathscr{I}_t = \{\boldsymbol{I}_t^i\}_{i=1}^4 \in \mathbb{R}^{H\times W\times C}$ as visual input, comprising three external viewpoints (wrist camera 1, 2 and side camera 3) and an active viewpoint (camera 4). We use pretrained DINOv2 [28] ViT as visual encoder for each $\boldsymbol{I}_t^i$, which produces $16\times 22$ tokens as scene representation. The proprioceptive state $\boldsymbol{p} \in \mathbb{R}^{19}$, includes the end-effector poses (position and quaternion) of two arm ($\in \mathbb{R}^{7\times 2}$) with two gripper ($\in \mathbb{R}^2$), 2-DoF gimbal angles ($\in \mathbb{R}^2$), and camera zoom (1 scalar). The policy outputs a sequence of future actions $\boldsymbol{a}_t = \{a_{t+1}, \ldots, a_{t+n}\} \in \mathbb{R}^{n\times 19}$, where each $a_{t+k}$ comprises two arm end-effector poses with gripper widths, gimbal angles, and zoom setting (all expressed in the world frame).

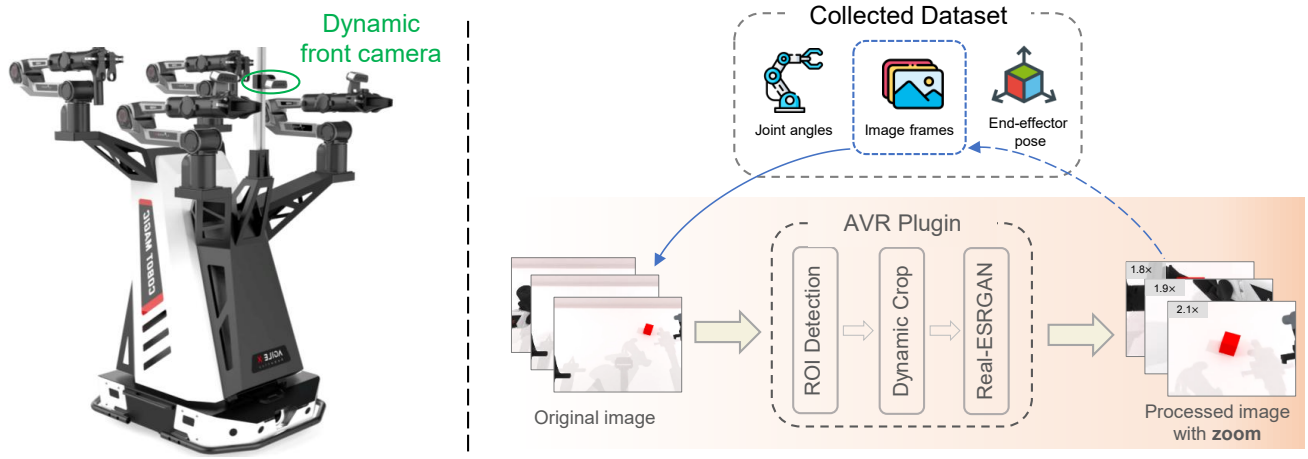


Fig. 3: AVR Plugin in simulation. We take the front camera view from RoboTwin collected dataset as input, get detailed observation by ROI detection, aspect-ratio crop with zoom, and Real-ESRGAN super reconstruction. The processed images with zoom are appended to dataset as additional detailed observation for policy.

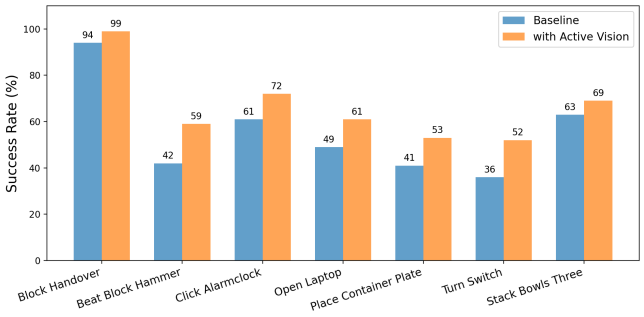


Fig. 4: Comparison of tasks success rates between baseline and Active Vision in simulation, which indicate that detail observation help policies better execute manipulation.

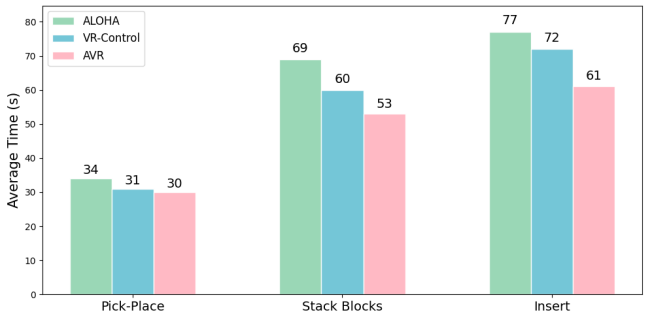


Fig. 5: Average task completion time under different teleoperation settings. Across representative tasks, operators using AVR complete demonstrations in less time compared to ALOHA and VR-control, indicating fewer failed attempts and more efficient one-shot executions.

## 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.

TABLE I: Trials needed to reach 50 successful demonstrations per task (successes/total).

| **Task** | ALOHA | VR-control | AVR |
|---|---|---|---|
| Pick-place | 50/50 | 50/50 | 50/50 |
| Dish scrubbing | 50/52 | 50/53 | 50/50 |
| Fold cloth | 50/51 | 50/53 | 50/50 |
| Place cup on rack | 50/50 | 50/52 | 50/50 |
| Block stacking | 50/56 | 50/60 | 50/52 |
| Grasp chewing gum | 50/61 | 50/57 | 50/55 |
| Insert screwdriver | 50/69 | 50/58 | **50/53** |

### A. RoboTwin-based Simulation Evaluation

To evaluate whether detailed observation benefits policy performance, we extend RoboTwin [13], a simulation data collection platform, with an active vision module (named AVR Plugin) and conduct diverse manipulation tasks. Shown in Fig. 3, we perform offline processing on collected dataset containing RGB images, joint angles, and end-effector poses. By extracting the front camera view, which covered the entire manipulation workspace, as AVR Plugin's input. First, we applied task-conditioned ROI detection on target objects to simulate **dynamic camera viewpoints**; Then we executed an aspect-ratio-preserving crop and computed the relative zoom ratio to simulate **zoom variation**. To match the expected detail level after zooming, we applied Real-ESRGAN [29], a state-of-the-art super resolution algorithm, to super-resolve the cropped region back to the original image resolution. The processed images, together with their corresponding zoom ratio, were integrated as additional details observation.

We deployed 50 expert demonstration for each task,

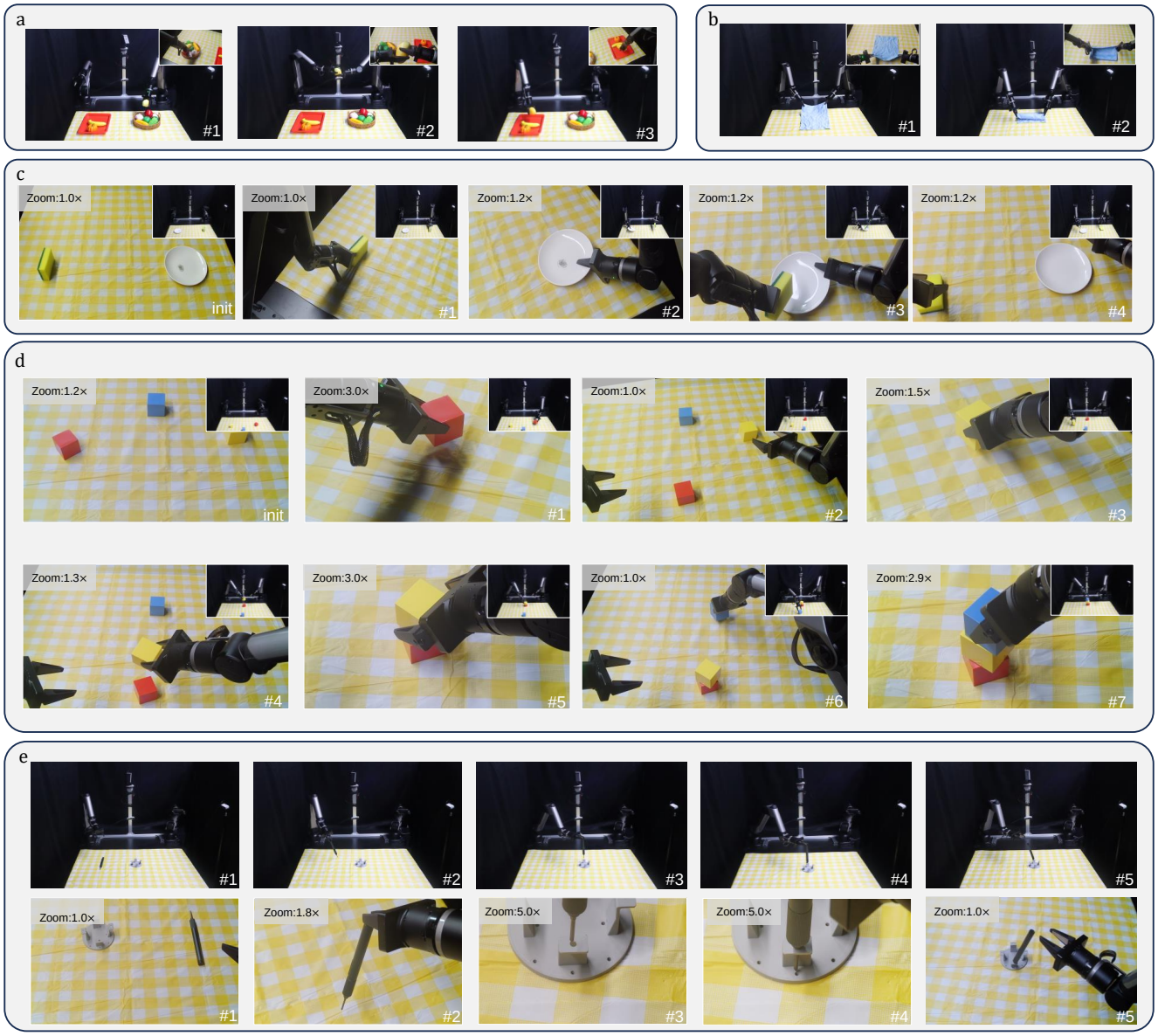


Fig. 6: Deployment of various manipulation tasks. (a) Pick-and-place of objects with varied shapes. (b) Folding fabric with coordinated bimanual manipulation. (c) Dish scrubbing with a controlled wiping motion. (d) Block stacking requiring precise alignment. (e) Inserting a screwdriver tip into a hole 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.

trained the model based on Diffusion Policy, and evaluate their performance in simulation. As shown in Fig. 4, our approach yielded 5%-17% increase across baseline, indicating that detail-aware observations help policies better understand and execute manipulation goals.

### *B. Real Robot Performance*

We designed a suite of real-world bimanual manipulation tasks on our hardware platform to assess the effect of active vision on overall performance, espcially precision manipulation. Representative real-robot deployments are illustrated in Fig. 6: (a) grasp a mango, perform a handover, and place it on a plate; (b) fold a towel once; (c) scrub a stained round dish; (d) sequentially grasp and stack three cubes (edge length 5 cm); and (e) grasp a small screwdriver (shank diameter 1 cm, head 0.5 cm) and accurately insert it into a fixed hole of 0.75 cm diameter. We further compare data collection efficiency and reliability across different teleoperation settings. Fig. 5 reports the average completion time across representative tasks, where operators using AVR complete trajectories more efficiently with fewer erroneous attempts, resulting in more consistent one-shot executions. Table I summarizes the number of failed trials during data collection, showing that AVR significantly reduces failures compared

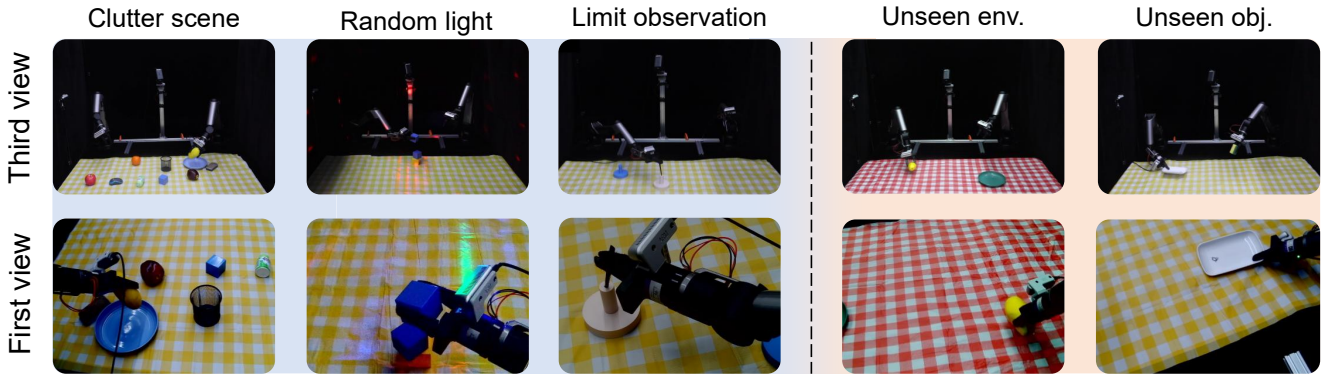


Fig. 7: Extended experiments. We assess policy robustness and generalization on the base tasks under four conditions: cluttered scenes and random lighting (left), and transfer to unseen environments and unseen objects (right), with third and first real-time view.

to conventional settings. This not only indicates improved data quality, but also demonstrates that even novice operators can more easily collect reliable demonstrations with our framework. For each task, we collected 50 demonstrations under our hardware platform and trained policies using a diffusion policy framework to evaluate performance across conditions. Augmenting both data collection and deployment with AVR yields higher success rates on most tasks, substantially improving manipulation performance.

We further analyzed the differential impact of viewpoint and focal-length adjustments on task success. Prior works [30], [31] indicate that optimizing camera viewpoint improves grasping in clutter, whereas increasing magnification/ zoom benefits fine motor tasks (sometimes at the cost of completion time). Our results align with these trends: for typical pick-place tasks (e.g., handover, towel folding, dish scrubbing), maintaining target visibility demands substantial top-camera viewpoint changes, while only minor zoom ($\leq 2\times$) is required; consequently, focal-length variation offers limited marginal gains. In contrast, for precision tasks (e.g., three-block stacking and small-hole insertion), viewpoint changes alone are insufficient to localize boundaries or apertures accurately. Adding dynamic focal length to obtain high-resolution local detail markedly improves alignment and positioning at critical phases, leading to higher success rates.

In summary, coarse pick-and-place tasks benefit primarily from viewpoint control, while precision manipulation benefits primarily from zoom. When the policy is equipped to actively acquire high-resolution detail at key moments, together with adequate viewpoint coverage, reliability and robustness are significantly improved on complex, high-precision tasks.

### C. Extended Analysis

In this section, we evaluate our method through multiple quantitative experiments, including ablation study, robustness assessment, and generalization validation. Some experimental results are presented in Fig. 7, and more deploy videos can be found on our `Project Page`.

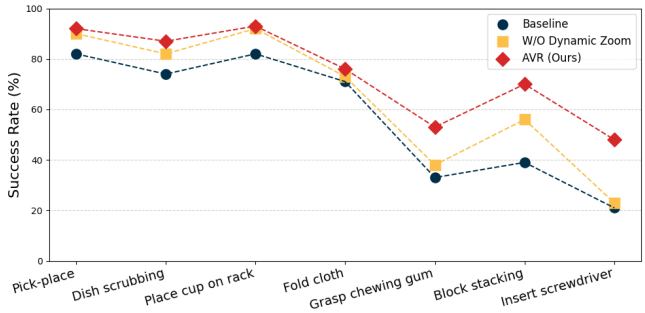


Fig. 8: Ablation study for active vision. We compare static view, dynamic viewpoint only, and our AVR (dynamic viewpoint with focal length) on several tasks. AVR consistently achieves higher success rates, especially on fine-grained tasks.

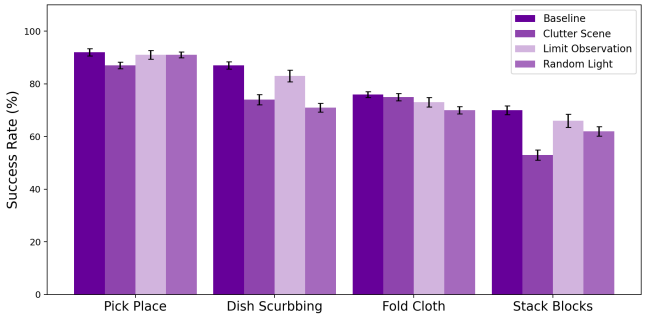


Fig. 9: Robustness under perturbations. We choose four typical manipulation tasks with three disturbance: cluttered scene, random light and limit observation. Result shows that occlusion induces the largest degradation on visually demanding tasks, highlighting the importance of detail-aware observation for robustness.

**Ablation Study**: To quantify the contribution of our AVR framework, we conduct ablations on visual observation, comparing three settings: (i) no active vision, (ii) dynamic viewpoint only (traditional active vision method), and (iii) dynamic viewpoint and focal length (our method). Fig. 8 shows that with additional detailed observation, success rates rise across all tasks, notably on fine-grained manipulations.

**Robustness**: Robust control under realistic perturbations is a core criterion of policy competence. We evaluate four manipulation tasks under three disturbance families: (i) limited observability via viewpoint occlusion, (ii) cluttered scene, and (iii) illumination disturbance, while randomizing object start poses. Fig. 9 summarizes the success rates (mean ± 95% CI). Tasks that rely more on visual observation (e.g., cloth folding) show more degradation under occlusion, highlighting the importance of detailed observation in maintaining robustness.

**Generalization**: A policy's ability to operate in unseen environments and with unseen objects is key to distinguishing task understanding from mere trajectory cloning. We evaluate generalization on multiple tasks under two settings: (i) unseen environments, where we vary scene layout and tabletop background; and (ii) unseen objects, where targets are replaced by new instances of the same category. Our method maintains higher success in unseen settings and shows smaller from seen to unseen, demonstrating robust cross-environment and cross-instance generalization enabled by viewpoint-and-focal length active vision.

## 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% - 17%, with more than **25%** increases in precision for precision tasks, significantly outperforming conventional imitation learning methods. Extended experiments confirm that our AVR framework, by enabling detailed observation and learning, yields substantial policy gains even when only limited data are available. The resultant policies demonstrate enhanced robustness under occlusion, clutter, and lighting disturbances, exhibiting smaller performance declines. Furthermore, they maintain higher success rates and smaller performance gaps when deployed in unseen environments and on novel objects, showcasing superior cross-environment and cross-instance generalization and underscoring the critical role of detail-centric observation.

Future work will address several key areas. First, we will improve data collection efficiency by refining teleoperation alternatives – e.g., using AR glasses and data gloves – to capture human active perception without robot teleoperation. Second, we will enhance viewpoint control by upgrading the gimbal mechanism and the VR-to-camera mapping to better track head motion, and by integrating additional sensors (e.g., wrist-mounted cameras) for richer observations. Finally, on policy learning, we will introduce instruction-conditioned (language-conditioned) policies to guide active perception.

## REFERENCES


[1] T. Osa, J. Pajarinen, G. Neumann, *et al.*, "An Algorithmic Perspective on Imitation Learning," *Foundations and Trends® in Robotics*, vol. 7, pp. 1–179, 2018.

[2] S. Ross, G. Gordon, and D. Bagnell, "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 15, pp. 627–635, 2011.

[3] J. Ho and S. Ermon, "Generative Adversarial Imitation Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[4] C. Chi, Z. Xu, S. Feng, *et al.*, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," *The International Journal of Robotics Research (IJRR)*, pp. 02783649241273668, 2023.

[5] Y. Mu, T. Chen, S. Peng, *et al.*, "RoboTwin: Dual-Arm Robot Benchmark with Generative Digital Twins (Early Version)," arXiv, 2024.

[6] X. Cheng, J. Li, S. Yang, *et al.*, "Open-TeleVision: Teleoperation with Immersive Active Visual Feedback," arXiv, 2024.

[7] I. Chuang, A. Lee, D. Gao, *et al.*, "Active Vision Might Be All You Need: Exploring Active Vision in Bimanual Robotic Manipulation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7952-7959, 2025.

[8] H. Xiong, X. Xu, J. Wu, *et al.*, "Vision in Action: Learning Active Perception from Human Demonstrations," arXiv, 2025.

[9] Q. Bu, J. Cai, L. Chen, *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," arXiv, 2025.

[10] M. Mittal, C. Yu, Q. Yu, *et al.*, "Orbit: A Unified Simulation Framework for Interactive Robot Learning Environments," *IEEE Robotics and Automation Letters*, vol. 8, pp. 3740–3747, 2023.

[11] Y. Wang, Z. Xian, F. Chen, *et al.*, "RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation," arXiv, 2023.

[12] A. Khazatsky, K. Pertsch, S. Nair, *et al.*, "DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset," arXiv, 2024.

[13] T. Chen, Z. Chen, B. Chen, *et al.*, "RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation," arXiv, 2025.

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

[15] G. Wang, H. Li, S. Zhang, *et al.*, "Observe Then Act: Asynchronous Active Vision-Action Model for Robotic Manipulation," *IEEE Robotics and Automation Letters*, vol. 10, pp. 3422-3429, 2025.

[16] D. Rakita, B. Mutlu, and M. Gleicher, "Remote Telemanipulation with Adapting Viewpoints in Visually Complex Environments," *Robotics: Science and Systems XV*, 2019.

[17] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation," in *Conference on Robot Learning (CoRL)*, pp. 785–799, 2023.

[18] J. Shang and M. S. Ryoo, "Active Vision Reinforcement Learning under Limited Visual Observability," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol.29, pp.10316-10338, 2023.

[19] J. Fu, H. Zheng, and T. Mei, "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4438–4446, 2017.

[20] Z. Yang, T. Luo, D. Wang, *et al.*, "Learning to Navigate for Fine-grained Classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 420-435, 2018.

[21] Z. Wang, Y. Yin, J. Shi, *et al.*, "Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 267-275, 2017.

[22] Q. Feng, X. Xu, and Z. Wang, "Deep learning-based small object detection: A survey," *Mathematical Biosciences and Engineering*, vol. 20, pp. 6551–6590, 2023.

[23] I. Chuang, A. Lee, D. Gao, *et al.*, "Look, Focus, Act: Efficient and Robust Robot Learning via Human Gaze and Foveated Vision Transformers," arXiv, 2025.

[24] "Galaxea A1."`https://github.com/userguide-galaxea/A1_SDK`. Accessed: Sep. 14, 2025.

[25] Nakanishi, Jun, Itadera, Shunki, Aoyama, Tadayoshi, *et al.*, "Towards the development of an intuitive teleoperation system for human support robot using a VR device." *Advanced Robotics*, vol. 34, pp. 1239-1253, 2020.

[26] G. Casiez, N. Roussel, and D. Vogel, "1€ Filter: A Simple Speed-based Low-pass Filter for Noisy Input in Interactive Systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 2527–2530, 2012.

[27] M. Schwarz and S. Behnke, "Low-Latency Immersive 6D Televisualization with Spherical Rendering," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 320-325, 2020.

[28] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," arXiv, 2023.

[29] X. Wang, L. Xie, C. Dong, *et al.*, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1905-1914, 2021.

[30] D. Morrison, P. Corke, and J. Leitner, "Multi-View Picking: Next-best-view Reaching for Improved Grasping in Clutter," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8762-8768, 2019.

[31] T.-C. Lin, A. U. Krishnan, and Z. Li, "erception and Action Augmentation for Teleoperation Assistance in Freeform Telemanipulation," *ACM Transactions on Human-Robot Interaction*, vol. 13, pp. 1–40, 2024.