ActiveUMI: Universal Manipulation Interface with Active Perception for In-The-Wild Robot Learning

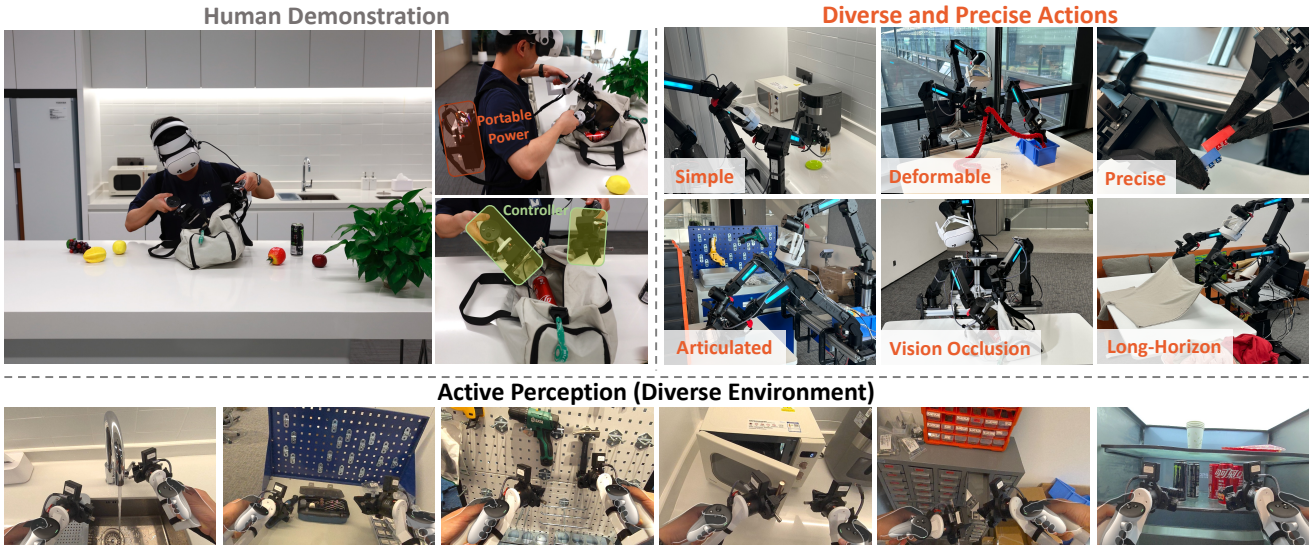


Fig. 1: **Universal Manipulation Interface with Active Perception (ActiveUMI)** is a portable, low-cost data collection framework for transferring diverse, in-the-wild human demonstrations into effective visuomotor policies. The core of our method is to empower this system with active perception, which allows the robot to control its viewpoint. This capability is critical for completing long-horizon tasks, overcoming visual occlusions, and performing actions that require high precision.

***Abstract*— We present ActiveUMI, a framework for a data collection system that transfers in-the-wild human demonstrations to robots capable of complex bimanual manipulation. ActiveUMI couples a portable VR teleoperation kit with sensorized controllers that mirror the robot's end-effectors, bridging human-robot kinematics via precise pose alignment. To ensure mobility and data quality, we introduce several key techniques, including immersive 3D model rendering, a self-contained wearable computer, and efficient calibration methods. ActiveUMI's defining feature is its capture of active, egocentric perception. By recording an operator's deliberate head movements via a head-mounted display, our system learns the crucial link between visual attention and manipulation. We evaluate ActiveUMI on six challenging bimanual tasks. Policies trained exclusively on ActiveUMI data achieve an average success rate of 70% on in-distribution tasks and demonstrate strong generalization, retaining a 56% success rate when tested on novel objects and in new environments. Our results demonstrate that portable data collection systems, when coupled with learned active perception, provide an effective and scalable pathway toward creating generalizable and highly capable real-world robot policies.**

I. Introduction

Robot foundation models promise generalist policies but are currently constrained by the scale and alignment of available robot data relative to web-scale corpora. A central challenge is therefore scaling data collection while preserving embodiment fidelity. Prevailing sources—in-lab teleoperation, human videos, and simulation—each have limitations: teleoperation is costly to scale; human videos [1]–[5] incur a cross-embodiment gap (human to robot); and simulation suffers a sim-to-real gap (physics to hardware [6]).

A promising middle ground is sensorized hand-held interfaces (e.g., grippers, dexterous-hand devices) that capture action-aligned trajectories. Yet most current interfaces overlook active, egocentric perception: humans move their heads to manage occlusion and gather context, while existing rigs rely primarily on wrist-mounted cameras. Even with a wide field-of-view, an end-effector–centric view underserves long-horizon tasks and fine manipulation and misaligns with platforms that use head-mounted cameras. These observations motivate data-collection and policy-learning pipelines that couple head-ego sensing with wrist-eye control, enabling viewpoint selection as part of the task and improving transfer to real robots.

To this end, we propose ActiveUMI, a universal manipulation interface with active perception for in-the-wild robot policy learning. Our approach is built on two core principles for scalable data collection: (i) the system must tightly align the robot's embodiment with natural human

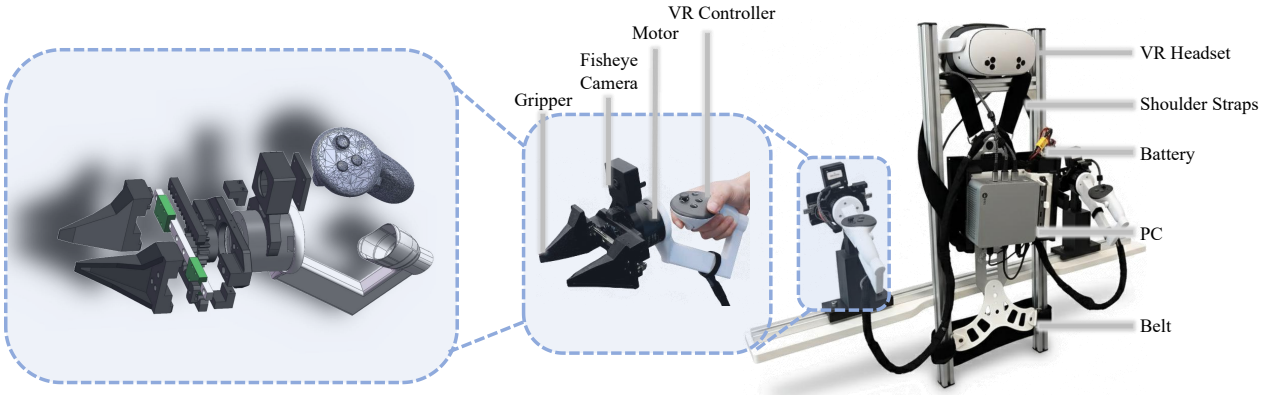


Fig. 2: **Overview of ActiveUMI Hardware.** A VR headset with custom controllers designed to replicate the structure of the robot's grippers. A portable backpack that holds a battery and a PC for self-contained operation.

movement, and (ii) it must enable active perception to expose the right sensory information at the right time. Our system addresses these needs with a specially designed, portable VR teleoperation kit. We developed a hardware architecture that allows the target robot's own custom grippers to be mounted directly onto the VR controllers, mirroring the end-effectors precisely. The entire system is self-contained in a backpack, and we implement several calibration techniques to ensure consistent, high-quality data collection in diverse real-world environments. To enable active perception, we map the operator's head movements to a movable robotics arm with a head-mounted camera. This allows the learned policy to control its own viewpoint, actively seeking out information to solve complex, long-horizon, or visually occluded tasks that are challenging for systems with only static or wrist-mounted cameras.

We evaluate ActiveUMI on six challenging, real-robot bimanual tasks that combine precise hand–object interactions with long-horizon manipulation using only the egocentric head camera and wrist proprioception available to the robot platform. By training policies trained purely on ActiveUMI demonstrations, they attain an average 70% success rate on all tasks. Relative to non-active perception counterparts (i.e., policies trained from wrist-centric views or static third-person cameras), ActiveUMI improves average success by 44% and 38%, respectively. Furthermore, when evaluated with novel objects and scenes, learned policies retain 56% of the average success rate, indicating a meaningful generalization from in-wild data.

II. Related Work

Data collection is a central pillar of modern deep learning, especially in the era of large models with massive numbers of learnable parameters. In robotics, the development of robot foundation models [7]–[11], such as Vision-Language-Action (VLA) models [12]–[15], has recently garnered significant attention. A critical prerequisite for training a robust and useful robot foundation model is the collection of massive datasets. However, the scale of today's robotics data is only a small fraction of that used for training large language models. Several approaches aim to alleviate this data scarcity problem, including designing user-friendly teleoperation systems [3], [16]–[20], leveraging large-scale simulation data [6], and repurposing human videos [4], [5], [21]–[23]. However, each has significant drawbacks: teleoperation is expensive and difficult to scale, while both simulation and human videos suffer from significant reality and embodiment gaps, respectively.

To overcome the scaling limitations of in-lab setups, research has explored collecting data "in-the-wild". One common source is using human demonstrations. DexCap [24] uses a wearable glove to capture precise wrist and fingertip poses for dexterous tasks. AirExo [25], [26] leverages low-cost hardware with direct kinematic mapping for arm manipulation. DoGlove [27] uses a low-cost, precise, and haptic force feedback glove system for teleoperation and manipulation. Dexop [28] uses a passive hand exoskeleton designed to maximize human ability to collect rich sensory data for diverse dexterous manipulation tasks in natural environments. NuEXO [29] designs a portable exoskeleton hardware to do both teleoperation and collect humanoid data. The Universal Manipulation Interface (UMI) [30] is the most related work to us. The UMI introduced a simple handheld controller for collecting bimanual data at scale, which DexUMI [31] later extended to dexterous hands with similar concepts. FastUMI [32] uses a substantial redesign of the UMI system that addresses these challenges by enabling rapid deployment via adding an extra camera on top of the UMI gripper. However, a common limitation among these systems is their primary reliance on wrist-mounted cameras for perception. Because these cameras move with the arm, their viewpoints are constrained by manipulation needs rather than by perceptual objectives, which makes it difficult to handle tasks with visual occlusions or long-horizon goals. Our work, ActiveUMI, is designed to bridge this specific gap. The ActiveUMI's core contribution is the integration of active, egocentric perception by explicitly tracking the operator's head movements via a VR headset. This allows the learned policy to actively control its own viewpoint—a capability that is critical for overcoming occlusions and successfully completing complex tasks.

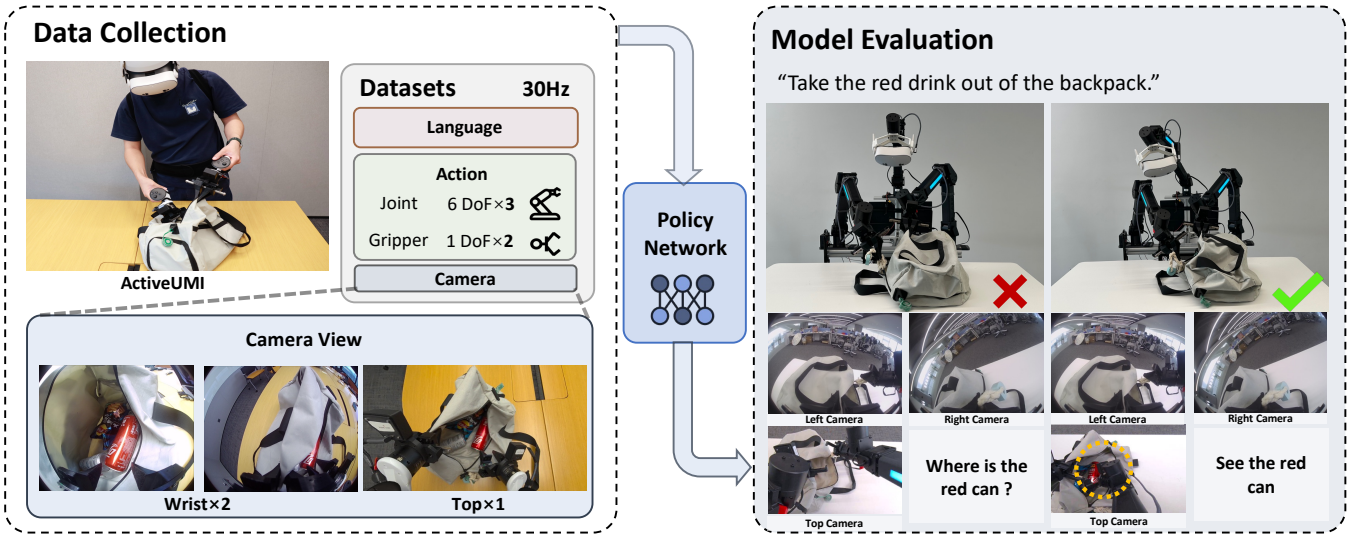


Fig. 3: **Overview of ActiveUMI.** The left side of the figure illustrates our data collection process and the detailed dataset configuration. The training data from in-the-wild data collected by ActiveUMI. The right side of the figure shows the model deployment and inference process.

III. METHODOLOGY

This section introduces ActiveUMI, a high-mobility framework designed for large-scale, in-the-wild robot learning. We will first provide an overview of the data collection system, then delve into the core concept of active perception, and conclude with the calibration methods that ensure high-quality data.

A. Data Collection System for ActiveUMI

Data collection is a central pillar of modern deep learning, especially in the era of large models with massive numbers of learnable parameters. In robotics, the development of robot foundation models [7]–[9], such as Vision-Language-Action (VLA) models [12], [33], has recently garnered significant attention. A critical prerequisite for training a robust and useful robot foundation model is the collection of massive datasets. However, the scale of today's robotics data is only a small fraction of that used for training large language models. Several approaches aim to alleviate this data scarcity problem, including designing user-friendly teleoperation systems, leveraging large-scale simulation data , and repurposing human videos. However, each has significant drawbacks: teleoperation is expensive and difficult to scale, while both simulation and human videos suffer from significant reality and embodiment gaps, respectively.

The design of ActiveUMI facilitates an intuitive and efficient process for high-quality data collection while extending the operational boundaries from constrained laboratory settings to diverse, "in-the-wild" environments. To this end, we have developed a low-cost, high-precision hardware system based on consumer-grade VR equipment, with its overall architecture depicted in Figure 2.

VR Gripper Controller. Our VR controller is a modified version of the commercial Meta Quest 3s controller, leveraged for its inherent capability for synchronous, low-latency, and high-precision six-degrees-of-freedom (6-DoF) pose tracking. This is accomplished via the headset's sophisticated inside-out tracking system. The headset's onboard cameras continuously triangulate the controller's pose in real-time by tracking a unique pattern of integrated infrared (IR) LEDs. By obtaining the 6-DoF pose data, we can concurrently resolve both the controller's translational position (x,y,z) and its rotational orientation (roll, pitch, yaw) within the captured volume. Consequently, by rigidly mounting this controller onto our target robot, its pose becomes directly representative of the robot's pose. A detailed analysis of the measurement error is provided in Section IV-E.

Our approach offers greater hardware flexibility compared to systems like UMI, which are often built around a specific, non-interchangeable gripper. We can adapt our system by simply mounting a modified Meta Quest controller onto the target robot's existing end-effector.

Gripper Actuation. We integrate a micro-motor directly onto the controller to drive the open-close motion of the gripper. This allows an operator to control the robot's grasp intuitively. A key advantage of our design is that it's non-invasive; instead of replacing the robot's "vanilla" gripper, we attach an identical copy to the operator's controller for data collection. This ensures our system can be deployed on a wide range of stock robots with minimal modification.

To enrich the data stream, we augment each controller with a fisheye camera. This wrist-mounted camera is positioned to maximize its field of view, capturing comprehensive visual information of the robot's immediate operational environment. This provides the downstream policy model with rich visual context, and the resulting "wrist view" serves as a valuable complement to the first-person perspective from the head-mounted camera.

Head-Mounted Display (HMD). The Meta Quest3s HMD plays a dual, critical role within our framework. Firstly, it serves as a high-precision localization hub. Its

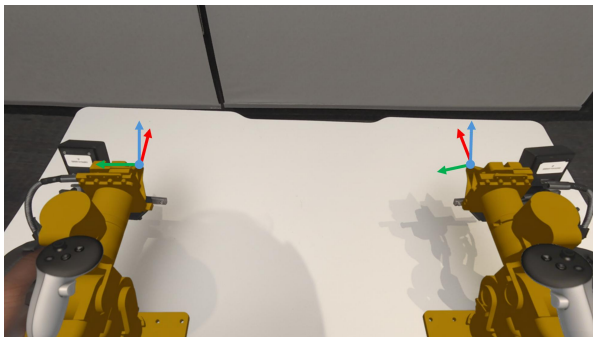


Fig. 4: **Immerse Data Collection.** Our system provides the operator with critical visual feedback by rendering the robot's arms in the VR environment.

robust SLAM system provides a stable and reliable world coordinate system, concurrently tracking the 6-DoF poses of both the operator's head and the controller. Secondly, the HMD's front-facing color cameras function as a dynamic, top camera, offering a global perspective that is intrinsically coupled with the operator's line of sight.

Wearable Device. To enable data collection in any environment, we utilize a compact, wearable computational unit consisting of a small computer worn on the operator's back. This self-contained design liberates the operator from a stationary workstation, allowing them to move freely and gather data across diverse settings.

Immerse Data Collection. To provide the operator with intuitive feedback, we render a 3D model of the robotic arms within the VR environment. These virtual arms are precisely aligned with the operator's hand-held controllers, which correspond to the robot's grippers. This setup allows the operator to clearly visualize the robot's movements in real-time during data collection. We visualize the rendered model in Figure 4.

B. Active Perception for Policy Learning

A key limitation of conventional UMI-style data collection is its reliance on wrist-mounted cameras. Because these cameras move with the robot's arms, their viewpoints are constrained by manipulation needs rather than guided by perceptual objectives. This makes it difficult for a trained policy to handle scenarios with visual occlusions, manipulate deformable objects, or perform tasks that require significant shifts in viewpoint.

ActiveUMI is designed to bridge this visual gap by enabling the robot to act with human-like flexibility in its head and camera control. To achieve this, we explicitly record the real-time 6-DoF pose of the operator's Head-Mounted Display (HMD) as an additional input to the policy. This allows the model to learn the crucial correlation between an operator's head movements (i.e., their visual attention) and their corresponding hand actions.

During deployment, the policy can then predict a 6-DoF pose for the robot's head, allowing it to actively mimic the operator's learned attention patterns. This predicted motion is executed by the robot's low-level controller, enabling the robot to dynamically adjust its viewpoint, overcome occlusions, and significantly enhance its performance on complex tasks.

C. Calibrating End-Effector for Precise Data Collection

The ActiveUMI system captures 6-DoF (Degrees of Freedom) pose data from three key points in the VR setup: the tips of the left and right controllers and the pose of the Quest 3 headset. During policy execution, these tracked points map one-to-one with the robot's two gripper tips and its head-mounted camera. All data is recorded in absolute coordinates relative to a unified world coordinate system that is established during an initial calibration phase. This ensures the reference frame remains consistent throughout the data collection session.

To ensure high-quality data alignment and maintain precision, we introduce three additional approaches to facilitate robust calibration.

In-Situ Environment Setup. To reset the 6-DoF zero-point, operators can press the 'B' button on the controller to reposition the base coordinate system. This feature enables data collection to start flexibly in any environment. The coordinate system's axes are rendered in real-time within the headset, allowing the operator to intuitively align the virtual reference frame with the physical workspace. This process ensures a consistent initial state for every data collection session.

Gripper Placeholder. To simplify calibration, we designed a physical placeholder that serves as a docking station for the VR controllers. This jig can be placed anywhere in the workspace to establish a consistent starting point. When the controllers are seated in the placeholder, their relative distance and pose are fixed to a predefined state. Pressing a designated button while the controllers are docked instantly calibrates the virtual coordinate system, aligning its origin and orientation with this known physical configuration.

Haptic Feedback for Zero-Point Position. To enhance the efficiency and convenience of zero-point calibration, we implemented a haptic feedback mechanism. Specifically, when a gripper moves within 3cm of the zero-point (the origin of the base coordinate system), the controller's motor generates a high-frequency vibration. This tactile cue alerts the operator that the gripper is approaching its base position. This mechanism allows users to confirm alignment without relying on numerical readouts, significantly improving the speed and efficiency of the calibration process.

By implementing the methods described above, we ensure that every data collection session begins from a precise and consistent initial pose. This guarantees an accurate one-to-one mapping between the operator's controls and the real robot's kinematics from the very start. Furthermore, these streamlined calibration procedures significantly reduce the operator's cognitive load. Ultimately, this user-centric design makes the data collection process more efficient and leads to higher-quality, more natural demonstrations, which is

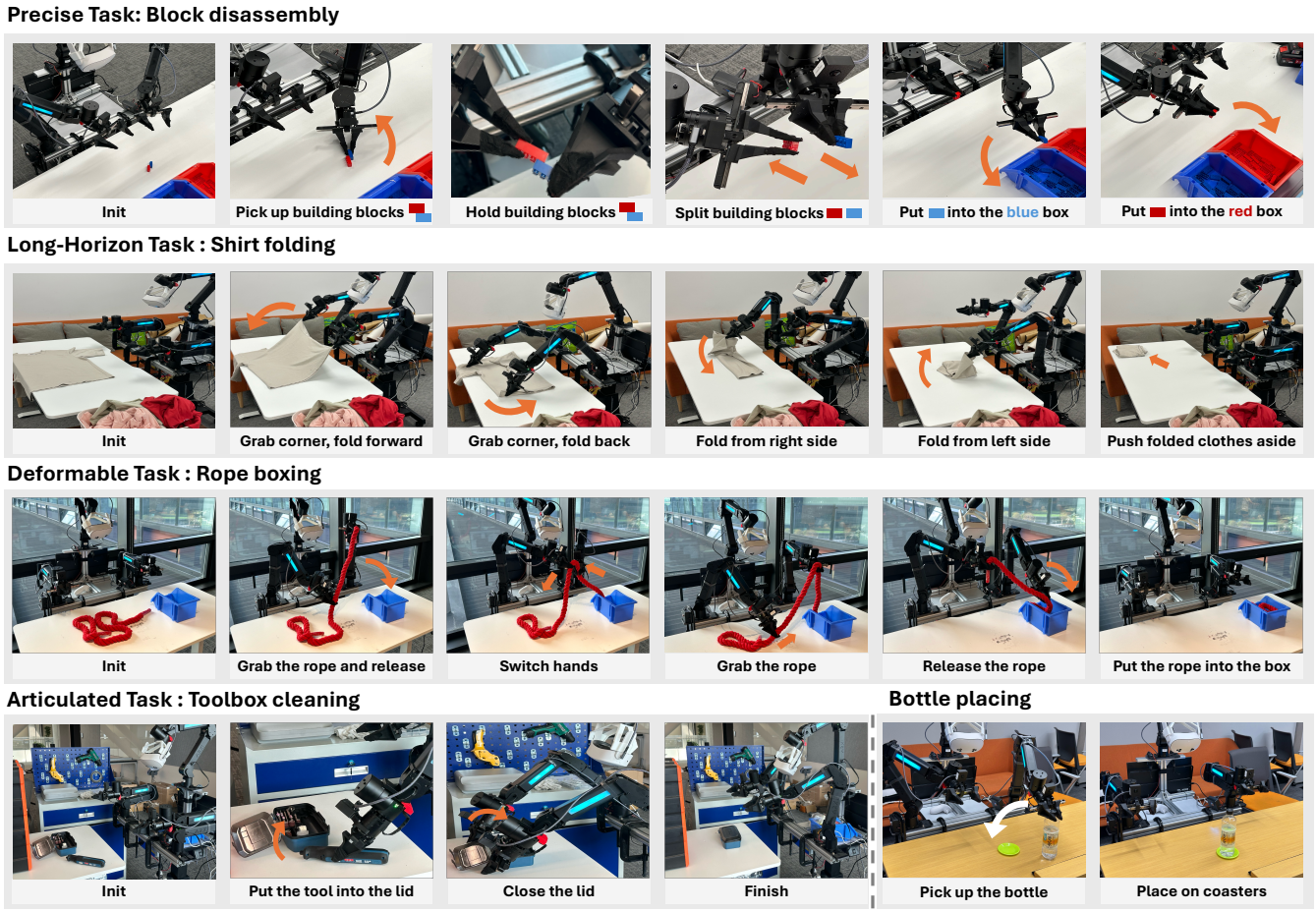


Fig. 5: **Evaluated Tasks.** We evaluated our approach on a diverse set of tasks, each requiring a different skill set: **Block disassembly** is a precision task where the robot must separate two small, interlocked blocks and then sort them into a box. **Shirt folding** is a deformable object manipulation task that demands accurate state recognition to correctly fold the cloth. **Rope boxing** is a long-horizon task where the robot must neatly place a long rope into a box. **Toolbox cleaning** is an articulated object manipulation task that requires the robot to close the lid. **Bottle placing** is a task designed to test the policy's robustness to large positional variations of the objects.

crucial for building a scalable framework for effective policy learning.

IV. Experiment

In this section, we will discuss the effectiveness of our proposed ActiveUMI. Specifically, we aim to investigate the following question:

- How important is the egocentric active perception for in-the-wild robot learning?
- What is the optimal strategy for utilizing ActiveUMI data to maximize end-to-end model performance?
- Can ActiveUMI data help the model generalize to new objects and scenes?

A. Implementation Details and Task Descriptions

Our real-world experiments are conducted on a testbed consisting of three 6-DoF ARX R5 robotic arms. Two arms, each equipped with a fisheye wrist-mounted camera, form a bimanual manipulation system. The third arm provides an active, mobile viewpoint, with its camera feed sourced from a human operator's VR headset to simulate an egocentric head camera. All sensor and robot data is collected at a frequency of 30Hz. For policy learning, we uses π_0, a state-of-the-art vision-language-action (VLA) model. For the fine-tuning stage, the model is subsequently fine-tuned for 50k iterations using a cosine learning rate scheduler. Unless otherwise stated, all experiments were conducted over 10 trials.

Our approach was evaluated on a diverse set of tasks, each designed to test a different robotic skill set:

- **Block disassembly**: A precision task requiring the robot to separate two small, interlocked blocks and sort them into a box.
- **Shirt folding**: A deformable object manipulation task demanding accurate state recognition to correctly fold the cloth.
- **Rope boxing**: A long-horizon task where the robot must neatly guide a long rope into a box.

TABLE I: We compare our active perception approach to two variants: a fixed top-down camera and a wrist-camera-only setup. The wrist-camera-only configuration corresponds to the UMI setting.

Camera View	Tasks (**In-Domain**)					
	Bottle placing	Rope boxing	Shirt folding	Block disassembly	Take Drink from Bag	Average
UMI	60%	20%	10%	0%	40%	26%
UMI w/ Fixed Head Camera	60%	40%	40%	20%	50%	42%
ActiveUMI	**90%**	**70%**	**80%**	**30%**	**80%**	**70%**

TABLE II: We compare our active perception approach to two variants in a new environment under the same task as Table I.

Camera View	Tasks (**New Environment**)					
	Bottle placing	Rope boxing	Shirt folding	Block disassembly	Take Drink from Bag	Average
UMI	30%	0%	0%	0%	0%	6%
UMI w/ Fixed Head Camera	30%	10%	20%	0%	20%	16%
ActiveUMI	**70%**	**50%**	**80%**	**30%**	**50%**	**56%**

- **Toolbox cleaning**: An articulated object manipulation task requiring the robot to operate a hinge to close the lid of a toolbox.
- **Bottle placing**: A task designed to test the policy's generalization and robustness to significant randomization in object positions.

We give an example for each task in Figure 5.

B. How Important is the Egocentric Active Perception?

A key feature of our proposed ActiveUMI framework is its use of active perception. This section investigates the impact of this component on model performance for complex manipulation tasks. Specifically, we compare the following three experimental setups:

- **Active Perception** (Our Method): The full ActiveUMI system, which includes a mobile head camera controlled by a dedicated 6-DoF arm (total 20-DoF).
- **Fixed Head Camera**: A baseline where the head camera is mounted in a static, top-down position, removing the active perception component (total 14-DoF).
- **Wrist-Camera-Only** (UMI Baseline): A second baseline where the head camera is removed entirely, leaving only the two fisheye wrist cameras. This configuration replicates the standard setup of UMI-style methods (total 14-DoF).

We use π_0 as the base model to train policies for all three configurations. For the wrist-camera-only baseline, we follow the official pi0 implementation and pad the visual tokens corresponding to the missing third-camera view. The experimental results, demonstrated in Table I, show that equipping the agent with active perception significantly outperforms both counterparts on all evaluated tasks. For instance, on the PourWater task, our method achieves a success rate 30% higher than the fixed top-down camera setup and 60% higher than the wrist-camera-only baseline.

We hypothesize two drivers of the improvements: (i) during in-the-wild data collection, demonstrators move their head and body; an active camera lets the policy compensate for this motion rather than treat it as observation noise; and (ii) active viewpoint selection enables the policy to acquire task-critical information (e.g., verifying a grasp) on demand. Finally, the fixed top-down camera reliably outperforms wrist-only, indicating that a third-person view adds complementary information for complex bimanual tasks.

TABLE III: **Data Mixing Ratio Experiments.** We conducted experiments on the shirt folding task to find the optimal data mixture for maximizing model performance.

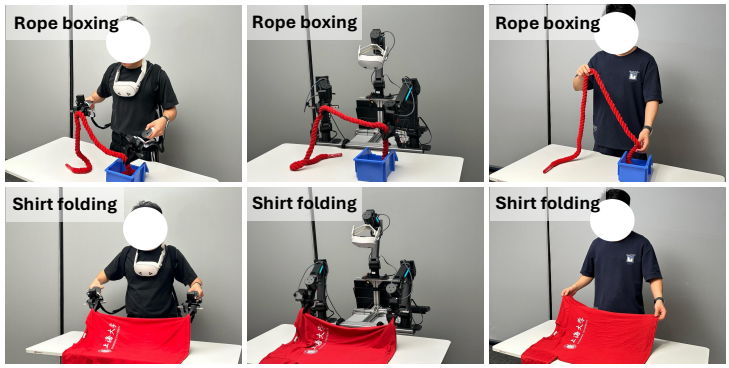
Teleoperated Data Ratio	10%	1%	0%
Avg. Success Rate	90%	95%	80%

C. Mixed Training with Teleoperated Data

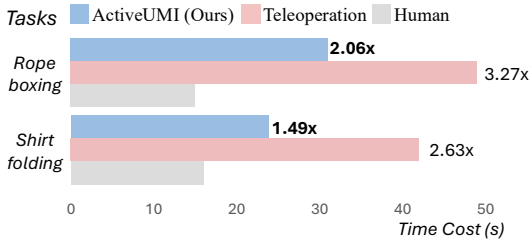
This section investigates the optimal strategy for using ActiveUMI data in policy training. To address the visual and embodiment gaps between human demonstrations and the robot, we evaluated a mixed-data approach on the complex, long-horizon shirt-folding task, conducting 20 trials for each experiment. Specifically, we compared three configurations: (1) training exclusively on ActiveUMI data, (2) mixing ActiveUMI data with 10% teleoperated data, and (3) mixing ActiveUMI data with only 1% teleoperated data.

The results, shown in Table III, indicate that adding teleoperated data improves performance. For instance, adding 10% teleoperated data increased the success rate from 80% to 90%. Interestingly, the optimal strategy was mixing in just 1% teleoperated data, which achieved a 95% success rate. This finding aligns with previous work showing that policies can be trained effectively by combining large-scale simulated data with a small amount of real-world demonstrations. This suggests that we can leverage large-scale, low-cost ActiveUMI data for effective model training, significantly lowering the cost of developing robot foundation models.

This demonstrates that ActiveUMI data is highly sample-efficient, requiring only a small fraction of real-world data to significantly boost and fine-tune the policy's performance. This conclusion aligns with findings from previous work, which have shown that policies can be effectively trained by mixing large-scale data with very few real-world teleoperated demonstrations.



(a) ActiveUMI Demonstration (b) Teleoperation (c) Human Bare Hand Demonstration



(d) Efficiency Comparison

	RPE(mm)
UMI	10.1
ActiveUMI (Ours)	**4.0**

(e) Relative Pose Error (RPE) Comparison

Fig. 6: **Data Collection Comparison.** (a)-(d) We utilize efficiency comparison among ActivateUMI, bare hand, and teleoperation in two tasks: rope boxing and shirt folding. ActivateUMI reaches an efficiency level between bare hand and teleoperation, and consistently outperforms teleoperation across both tasks.
(e) The comparison of relative pose error between UMI and ActiveUMI.

D. Generalization Capability of ActiveUMI for In-the-Wild Data Collection

A key indicator of a robust policy is its ability to generalize to novel objects and unseen scenes. To evaluate this capability, we tested the policies trained on ActiveUMI data in a new environment, performing the same set of tasks as the in-domain evaluation.

This experiment aims to determine if the skills learned, particularly active perception, can transfer to a different visual context. The results, presented in Table III, show that the policy trained with ActiveUMI demonstrates strong generalization capabilities. It achieves an average success rate of 56% in the new environment, retaining a significant portion of its in-domain performance.

Crucially, this performance significantly surpasses the baselines in the novel setting. The policy using a fixed head camera dropped to a 16% success rate, while the wrist-camera-only (UMI) baseline's performance fell to just 6%. This indicates that policies relying on more static or constrained viewpoints fail to adapt when the environment changes. In contrast, the ability to actively control its viewpoint allows the ActiveUMI policy to be more resilient to visual shifts. These findings validate that the "in-the-wild" data from ActiveUMI, enriched with active perception, produces policies that are not only effective but also generalizable.

E. Data Collection Throughput and Accuracy

Throughput. The previous section demonstrated that data collected by ActiveUMI is effective for training policies with active perception. A key advantage of our approach is its data collection efficiency. To evaluate this, we measured the time required to complete two long-horizon tasks—rope boxing and shirt folding—using three distinct methods: ActiveUMI, teleoperation of a real robot via a VR kit, and direct human demonstration.

As shown in Figure 6(d), ActiveUMI significantly speeds up data collection compared to teleoperation. For the rope boxing task, ActiveUMI was 2.06x slower than a direct human demonstration, while conventional teleoperation was 3.27x slower. Similarly, for shirt folding, ActiveUMI was 1.49x slower, compared to 2.63x for teleoperation. These results highlight that ActiveUMI provides a practical middle ground, retaining much of the efficiency of natural human motion while being substantially faster than conventional teleoperation.

Data Collection Accuracy. Having shown that ActiveUMI is effective as both a sole training data source and a supplement to teleoperated data, this section evaluates its collection accuracy. Specifically, we measure the error between the data collected by ActiveUMI and the actual trajectories replayed by the robot.

Specifically, we measure the Relative Pose Error (RPE). The experimental task was as follows: an operator, holding the ActiveUMI controller's gripper, placed the gripper at both ends of the tape measure, recording the nominal distance manually. The nominal distances started from 100cm and decreased in 10cm steps, sequentially collecting data for 100cm, 90cm, …, 10cm, for a total of 10 data points. During the experiment, the 6DoF pose sequences of the two grippers were recorded in real-time. We then analyzed the positioning accuracy of the ActiveUMI system based on this recorded data. Next, we entered the playback phase, where the saved pose sequences were precisely replicated on a real robot. At this point, we used the same tape measure to measure the actual distance between the inside of the two grippers, which was recorded as the *playback distance*. Using the nominal distance as the ground truth, we calculated the absolute error of the playback distance relative to the nominal distance:

$$\Delta L = \left| L_{\text{replay}} - L_{\text{measure}} \right|. \tag{1}$$

We further computed the relative error as:

$$RPE = \frac{\Delta L}{L_{\text{measure}}} \times 100\%. \tag{2}$$

We record the average RPE of ten trials and compares with the UMI. The experimental results are shown in Figure 6(e). We can observe that the RPE of UMI is 2.5x smaller than UMI. This low error is naturally comes from the advantange of the VR system, thus we obtain much better data quality and train train good policy network.

V. CONCLUSIONS

In conclusion, we identified a critical limitation in current robot data collection methods: the neglect of active, egocentric perception. While humans naturally move their heads to understand and interact with the world, most robot learning systems rely on action-centric, wrist-mounted cameras that limit performance on complex, long-horizon, or occluded tasks. To address this, we introduced ActiveUMI, a portable, in-the-wild data collection framework that couples high-fidelity embodiment alignment with learned viewpoint control. Our experiments demonstrate that this approach is highly effective. Policies trained exclusively on ActiveUMI data achieve a 70% success rate on a variety of challenging bimanual tasks. Crucially, our method significantly outperforms baselines that lack active perception, confirming that learning how to look is as important as learning what to do. The strong generalization performance on novel objects and scenes further validates the quality of in-the-wild data collected with this approach.

REFERENCES

[1] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," *arXiv preprint arXiv:2207.09450*, 2022.

[2] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen, G. Yang, J. Zhang, S. Yi, G. Shi, and X. Wang, "Humanoid policy - human policy," *arXiv preprint arXiv:2503.13441*, 2025.

[3] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[4] W. Ye, F. Liu, Z. Ding, Y. Gao, O. Rybkin, and P. Abbeel, "Video2policy: Scaling up manipulation tasks in simulation through internet videos," *arXiv preprint arXiv:2502.09886*, 2025.

[5] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen, "Video prediction policy: A generalist robot policy with predictive visual representations," *arXiv preprint arXiv:2412.14803*, 2024.

[6] A. Maddukuri, Z. Jiang, L. Y. Chen, S. Nasiriany, Y. Xie, Y. Fang, W. Huang, Z. Wang, Z. Xu, N. Chernyadev *et al.*, "Sim-and-real co-training: A simple recipe for vision-based robotic manipulation," *arXiv preprint arXiv:2503.24361*, 2025.

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

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

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

[10] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[11] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. J. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 16 923–16 930.

[12] K. Black, N. Brown, D. Driess *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[13] Physical Intelligence, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[14] NVIDIA, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[15] P. Ding, J. Ma, X. Tong, B. Zou, X. Luo, Y. Fan, T. Wang, H. Lu, P. Mo, J. Liu *et al.*, "Humanoid-vla: Towards universal humanoid control with visual integration," *arXiv preprint arXiv:2502.14795*, 2025.

[16] Y. Ze, Z. Chen, J. P. Araújo, Z. ang Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," 2025.

[17] W. Sun, L. Feng, B. Cao, Y. Liu, Y. Jin, and Z. Xie, "Ulc: A unified and fine-grained controller for humanoid loco-manipulation," *arXiv preprint arXiv:2507.06905*, 2025.

[18] Y. Li, Y. Lin, J. Cui, T. Liu, W. Liang, Y. Zhu, and S. Huang, "Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks," *arXiv preprint arXiv:2506.08931*, 2025.

[19] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," *arXiv preprint arXiv:2502.13013*, 2025.

[20] 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.

[21] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[22] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 13 226–13 233.

[23] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu *et al.*, "Egovla: Learning vision-language-action models from egocentric human videos," *arXiv preprint arXiv:2507.12440*, 2025.

[24] 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.

[25] 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," *arXiv preprint arXiv:2309.14975*, 2024.

[26] H. Fang, C. Wang, Y. Wang, J. Chen, S. Xia, J. Lv, Z. He, X. Yi, Y. Guo, X. Zhan, L. Yang, W. Wang, C. Lu, and H.-S. Fang, "Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons," *arXiv preprint arXiv:2503.03081*, 2025.

[27] H. Zhang, S. Hu, Z. Yuan, and H. Xu, "Doglove: Dexterous manipulation with a low-cost open-source haptic force feedback glove," *arXiv preprint arXiv:2502.07730*, 2025.

[28] H.-S. Fang, B. Romero, Y. Xie, A. Hu, B.-R. Huang, J. Alvarez, M. Kim, G. Margolis, K. Anbarasu, M. Tomizuka *et al.*, "Dexop: A device for robotic transfer of dexterous human manipulation," *arXiv preprint arXiv:2509.04441*, 2025.

[29] R. Zhong, C. Cheng, J. Xu, Y. Wei, C. Guo, D. Zhang, W. Dai, and H. Lu, "Nuexo: A wearable exoskeleton covering all upper limb rom for outdoor data collection and teleoperation of humanoid robots," *arXiv preprint arXiv:2503.10554*, 2025.

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

[31] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. Song, "Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation," *arXiv preprint arXiv:2505.21864*, 2025.

[32] K. Liu, C. Guan, Z. Jia, Z. Wu, X. Liu, T. Wang, S. Liang, P. Chen, P. Zhang, H. Song *et al.*, "Fastumi: A scalable and hardware-independent universal manipulation interface with dataset," *arXiv preprint arXiv:2409.19499*, 2024.

[33] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng, "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," *arXiv preprint arXiv:2502.05855*, 2025.