

# MV-UMI: Multi-View Scalable Interface for Cross-Embodiment Learning

Anonymous Submission

**Abstract**—Recent advances in imitation learning have shown great promise for developing robust robot manipulation policies from demonstrations. However, this promise is contingent on the availability of diverse, high-quality datasets, which are not only challenging and costly to collect but are often constrained to a specific robot embodiment. Portable handheld grippers have recently emerged as intuitive and scalable alternatives to traditional robotic teleoperation methods for data collection. However, their reliance solely on first-person view wrist-mounted cameras often creates limitations in capturing sufficient scene contexts. In this paper, we present MV-UMI (Multi-View Universal Manipulation Interface), a framework that integrates a third-person perspective with the egocentric camera to overcome this limitation. This integration mitigates domain shifts between human demonstration and robot deployment, preserving the cross-embodiment advantages of handheld data-collection devices. Our experimental results, including an ablation study, demonstrate that our MV-UMI framework improves performance in sub-tasks requiring broad scene understanding by approximately 47% across 3 tasks, confirming the effectiveness of our approach in expanding the range of feasible manipulation tasks that can be learned using handheld gripper systems, without compromising the cross-embodiment advantages inherent to such systems. Videos can be found here: <https://mv-umi.github.io>

## I. INTRODUCTION

Imitation Learning (IL) provides a compelling pathway toward acquiring general robot policies capable of performing long-horizon tasks across diverse environments. This approach, particularly through supervised methods like Behavioral Cloning (BC), enables robots to acquire complex behaviors by learning to imitate human-directed actions in response to observations. Recent advances in architectures that better model this mapping [1], [2], [3], [4], coupled with enhancements in embodiments and hardware integrations [5], [6], have made this route increasingly convincing.

Recent studies on data scaling laws in imitation learning [7] show that robot policy performance follows training scenario diversity, emphasizing the need for extensive and diverse data for robust policies. Data collection typically lies between two extremes. On one end, robot teleoperation enables the acquisition of high-quality, precise data with minimal embodiment discrepancies. However, this method is time-consuming and costly, as it requires an actively operated robot. On the other end, the internet is replete with videos of humans performing various tasks. However, substantial effort is required to establish structured explicit mappings between observed states and actions from these videos.

As a middle-ground, portable handheld grippers [8], [9], [10], [11], [12] have emerged as inexpensive and intuitive to use data collection devices. By relying exclusively on

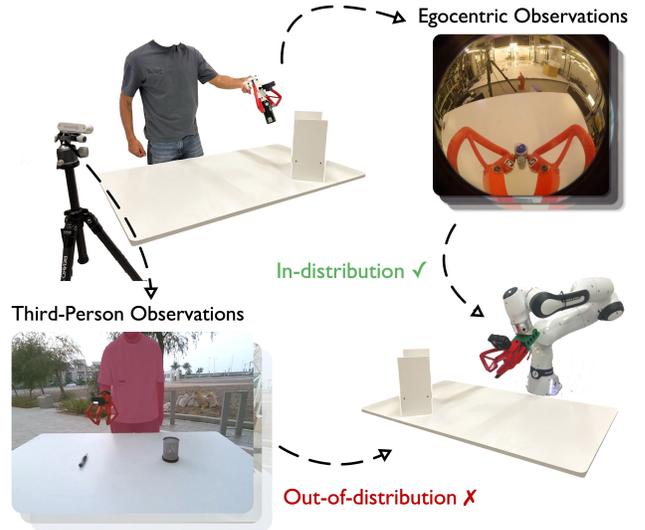


Fig. 1: Portable handheld gripper systems capture egocentric wrist-mounted views that remain consistent across human demonstrations and robot deployment. In contrast, third-person views are human-aligned during data collection but robot-aligned at deployment, creating an out-of-distribution gap that **MV-UMI** closes.

a wrist-mounted camera, they enable non-experts to record demonstrations without the need for a robotic manipulator. While this egocentric viewpoint minimizes visual discrepancies between training and deployment, resulting in cross-embodiment policies, it demands that the robot maintain a longer memory context to recall scene elements that move out of the constraining wrist view.

In this work, we propose a novel framework that augments the conventional wrist-mounted camera in handheld gripper systems with a third-person camera viewpoint, without incurring distributional shifts. We achieve this by performing real-time masking of the human demonstrator in the third-person video stream, effectively removing the operator’s presence from the training data. As a result, the model benefits from a broader view of the environment, while relying less on memory for scene remembrance. A side benefit we find of this masking is its removal of correlations between the demonstrator’s motions and the gripper’s actions, encouraging the policy to focus on task-relevant cues such as the manipulated objects rather than overfitting to human-specific signals. We also utilize a custom-made three-jaw gripper for some of the tasks that require greater dexterity in this work. This design allows for greater payload weight in comparison to other hand-held devices, at the cost of

its volume. Schematics and instructions to reproduce the hardware are open-sourced separately. The hardware aspect is not a key focus in this paper.

### Summary of Contributions:

- 1) **Multi-View Cross-Embodiment Framework:** MV-UMI fuses wrist-mounted and third-person views using SAM-2 segmentation and inpainting to eliminate domain shift, boosting performance in context-dependent tasks by 47%.
- 2) **End-to-End Open-Source System:** Complete pipeline, including hardware design, data collection, training code, and deployment tools, is publicly released to advance cross-embodiment manipulation research, <https://mv-umi.github.io>.

## II. RELATED WORKS

### A. Robot-Free Data-Collection Devices

Teleoperation has long been the standard approach for collecting demonstrations in robot learning, with systems ranging from general-purpose input devices such as phones and VR controllers [13], [14], [15] to more specialized hardware-based teacher–follower systems [5], [16], [17]. A common limitation across these approaches, however, is the requirement of having a physical robot available during data collection. In practice, this shifts the bottleneck from the availability of human demonstrators to the availability of robots, which remain costly and limited in access.

In recent years, portable handheld grippers have gained interest as a cost-effective alternative to teleoperation for collecting manipulation demonstrations. One of the earliest explorations in this was demonstrated by kitchen-inspired tongs [18], with motion capture markers to record kinematic trajectories. Subsequent systems [19], [20] advanced this work by incorporating a wrist-mounted camera on the device to capture a view that would align with the observation space during deployment on a robot. This allowed us to avoid reliance on motion capture markers in favor of directly extracting action-relevant features from recorded frames.

Building on these efforts, several handheld interfaces have been proposed to further scale data collection while improving data quality. In Dobb-E and RUM frameworks [8], [12], a low-cost “stick” device is used with an iPhone acting as a wrist-mounted sensor, capturing both RGB-D and motion data. Meanwhile, the UMI framework [9], [21] has become the go-to device for quick data collection, owing to its hardware-agnostic design and imitation learning pipeline. It utilizes a mounted GoPro camera to maximize the field-of-view and uses visual SLAM for posture estimation. Unlike Dobb-E’s approach however, this SLAM method requires pre-collection calibration and depends on sufficiently textured environments to maintain reliable tracking of pose. Complementing these, Fast-UMI [11] simplifies UMI deployment with a wrist-mounted tracking module, while Legato [10] introduces a handheld gripper that unifies action and observation spaces across embodiments.

TABLE I: Comparison of State-of-the-Art Cross-Embodiment Frameworks

Method	Viewpoints		Without Robot Teleoperation
	1st-Person	3rd-Person	
Dobb-E [8]	✓	✗	✓
UMI [9]	✓	✗	✓
Fast-UMI [11]	✓	✗	✓
Legato [10]	✓	✗	✓
Shadow [25]	✗	✓	✗
Mirage [26]	✗	✓	✗
<b>MV-UMI (Ours)</b>	✓	✓	✓

Beyond hand-held devices, works have explored other approaches for demonstration collection. DexCap [22] introduces a glove-based system that captures wrist and finger motions alongside egocentric RGB-D observations. In a complementary direction, AR2-D2 [23] removes the need for physical robots entirely by using iPhone’s AR application to overlays a virtual robot arm onto real-world scenes, allowing users to record demonstrations without the need of a physical robot.

While typical imitation learning setups [3] and datasets [24] that are collected using teleoperation use both third-person and egocentric viewpoints, current hand-held grippers rely solely on wrist-mounted cameras. This limits the ability of integrating other datasets during training, and necessitates that policies learn long temporal dependencies for objects that move out of view, particularly in multi-step tasks.

### B. Cross-Embodiment Learning

As we advance toward generalizable models that operate across diverse environments and platforms, enabling policies to transfer across different robot embodiments has become a critical area of exploration. The goal is to be able to transfer policies from one robot embodiment to another without the need for additional data collection.

Shadow [25] introduces a data editing technique that addresses this challenge using composite segmentation masks. During training, the source robot’s pixels are masked and replaced with a rendered segmentation mask of the target robot in the corresponding end-effector pose. This alignment ensures a consistent input distribution between training and evaluation, enabling policy transfer despite visual differences between robot embodiments. However, this technique assumes prior knowledge of the target embodiment during policy training, which limits its generalizability. In Mirage [26], image-inpainting techniques are used to replace the source robot with the target robot during evaluation. By “painting over” the source robot collected data with an inpainted image of the target, Mirage minimizes visual discrepancies without introducing additional deployment latency. This approach, however, also requires prior knowledge of the target embodiment during training.

A complementary line of work [27], [28] explores designing intermediate interfaces that abstract away embodiment-specific visual cues. Im2Flow2Act introduces an object-flow representation that tracks the motion of manipulated objects

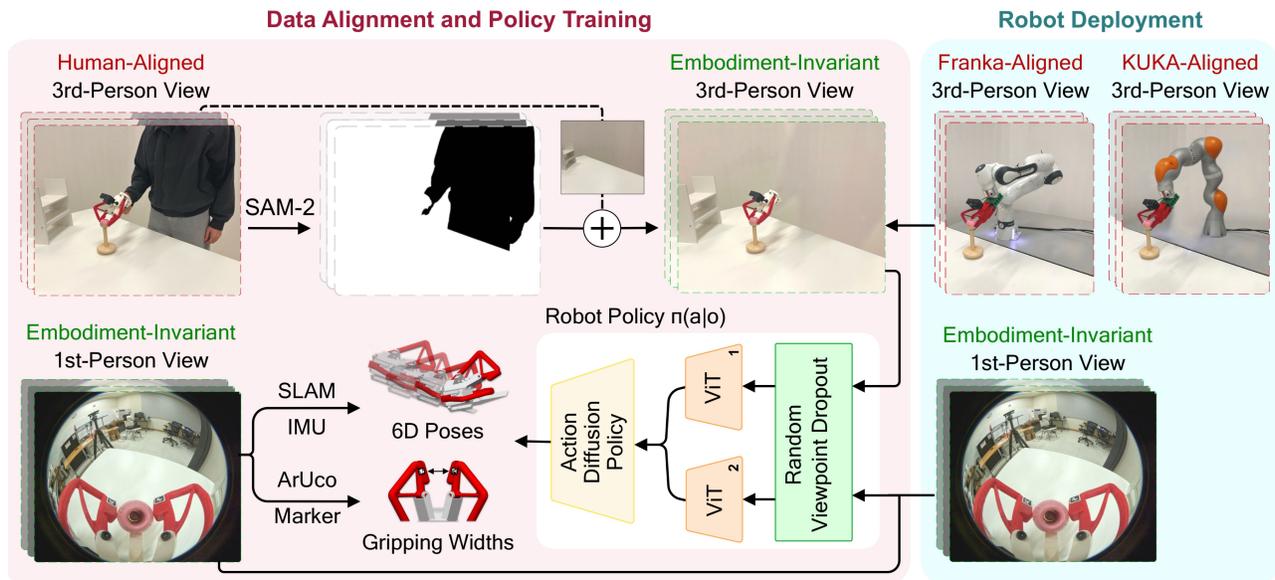


Fig. 2: MV-UMI data processing and deployment pipeline. During training, human demonstrations are captured from both egocentric (wrist-mounted) and third-person cameras, with SAM-2 segmentation removing the human demonstrator and background inpainting creating robot-compatible observations. At deployment, the trained policy processes real-time multi-view inputs for cross-embodiment policy transfer.

in the scene and uses this as the primary input to a diffusion-based policy for manipulation. Our MV-UMI framework shares the same goal of minimizing embodiment bias in demonstrations but takes a complementary approach. Rather than restricting the observation to only certain task-relevant signals, we instead remove embodiment-specific elements that should not be visible, preserving the rest of the scene context.

### III. MV-UMI FRAMEWORK

#### A. Problem Setup

We frame our task as a *cross-embodiment* learning problem in which the human demonstrator and the robot manipulator are considered distinct embodiments. Each recorded demonstration consists of:

- **Egocentric camera view:** A first-person image stream, denoted  $o_t^{\text{ego}}$ , recorded from the camera mounted on the handheld gripper device. This view is embodiment-invariant by nature, i.e., it remains the same in both the source and target settings.
- **Third-person camera view:**
  - During data collection, the overhead view is denoted as  $o_t^{\text{3rdH}}$  and captures the entire scene, including the human demonstrator.
  - During deployment, the overhead view is denoted as  $o_t^{\text{3rdR}}$  and also captures the scene but with the robot manipulator instead.

We do not formally calibrate the third-person camera placement across different environments. This design choice not only reduces setup effort and stays in touch with the in-the-wild data collection, but also introduces natural viewpoint variation that improves robustness during deployment.

- **Actions:** The handheld device pose  $T$  (i.e., end-effector position/orientation), along with the gripping width  $w$ , recorded at each time step.

We collect trajectories of the form  $\{(s_1, a_1), (s_2, a_2), \dots, (s_T, a_T)\}$ , where  $s_t = (o_t^{\text{ego}}, o_t^{\text{3rdH}})$  and  $a_t = (T_t, w_t)$  is the handheld device pose and gripper width. Our goal is to learn a policy  $\pi(a_t | s_t)$  that maps the state  $s_t$  to the corresponding actions  $a_t$ .

#### B. Data Collection and Preparation

Each recording session starts with the human operator scanning a code using the wrist-camera (done per environment, not episode). This initializes shared timestamps between the egocentric and third-person view cameras. Meanwhile, the third-person camera continuously records without interruption throughout the entire session. Once the raw videos are collected, we perform several offline processing steps as described below.

1) *Segmentation:* The third-person footage depicts the human demonstrator with the handheld gripper; thus, we remove the human from each overhead frame to eliminate correlations between the human’s movements and the robot’s intended actions.

Specifically, we apply Segment Anything v2 ( $SAM_2$ ) [29] on each overhead frame as in:

$$o_t^{\text{3rdH}} : o_t^{\text{3rd-mask}} = SAM_2(o_t^{\text{3rdH}}) \quad (1)$$

This process yields a binary mask of the human region for subsequent inpainting. To prompt  $SAM_2$ , multiple points are selected from the first frame of the first episode only: positive points that indicate the person and negative points indicating the gripper to ensure proper segmentation. The keypoints of that first frame are later propagated throughout



Fig. 3: Portion of environments used with MV-UMI demonstrating robustness in *in-the-wild* scenes.

the entire session, without the need for manual prompting in each episode.

2) *Inpainting with a Static Reference Frame*: To fill in the masked region, we use a background reference frame  $o_{\text{ref}}^{\text{bg}}$ . We then in-paint the human pixels by blending the background frame into the masked area via (2):

$$o_t^{\text{3rd-masked}} = \text{Inpaint}(o_t^{\text{3rdH}}, o_{\text{ref}}^{\text{bg}}, o_t^{\text{3rd-mask}}) \quad (2)$$

This produces overhead frames that appear as if no human was present, thereby mitigating distribution shift when transferring to a robot that sees only itself and the environment. As we show in our ablation study (Section IV), this human-removal step is critical for enabling successful policy transfer, with policies trained on unsegmented data not working robustly.

3) *Time Synchronization*: Although we initialize the camera streams via a code scan at the start, minor frame-to-frame offsets can still occur. Thus, each egocentric frame  $o_t^{\text{ego}}$  is matched to the processed overhead frame in the closest time  $o_{\tau(t)}^{\text{3rd-masked}}$ . This pairing ensures that both views capture the same scene context. We then store  $s_t$ , defined in (3) along with the corresponding action  $a_t$ :

$$s_t = \left( o_t^{\text{ego}}, o_{\tau(t)}^{\text{3rd-masked}} \right) \quad (3)$$

4) *Action Extraction*: In a similar fashion to the UMI [9], we extract actions directly from the egocentric .MP4 footage. In particular, visual SLAM using ORB-SLAM3 [30] is used together with GoPro IMU data to estimate the pose of the handheld device in each frame relative to the starting pose. The gripping width is determined by measuring the distance between two ArUco markers attached to the jaws of the handheld gripper. These pose and gripper-width estimates form the final action labels  $\{a_1, \dots, a_T\}$ .

### C. Policy Training

We parameterize our policy  $\pi$  via an action diffusion CNN-based architecture that predicts multi-step trajectories of end-effector poses and gripping widths. Rather than outputting a single action vector at each time step, the diffusion model infers a short-horizon sequence.

To improve robustness against potential artifacts such as imperfect inpainting and occlusions, we incorporate a random dropout strategy during training. This dropout probability  $p$  is exponentially reduced throughout the training, which encourages the network to avoid overfitting to a particular viewpoint or pattern. Random viewpoint dropouts and noise patches are also performed to make the model more robust

against occlusions and incomplete observations, as shown in Fig. 4.

### Algorithm 1 Policy Training with Random Dropout

- 1: Initialize parameters  $\theta$ , dropout rate  $p_0$ , decay rate  $\lambda$
- 2: **for** each iteration  $t = 1, \dots, T$  **do**
- 3: Sample raw frames  $\{o_t^{\text{ego}}, o_t^{\text{3rdH}}\}$  and ground-truth action  $a_t$
- 4: Apply SAM2 on  $o_t^{\text{3rdH}}$  to obtain  $o_t^{\text{3rd-masked}}$
- 5: **if** with probability  $(1 - p)$  **then**
- 6: Input both  $o_t^{\text{ego}}$  and  $o_t^{\text{3rd-masked}}$  to the policy
- 7: **else**
- 8: Randomly add noise patches to one view
- 9: **end if**
- 10: Compute predicted action:  $\hat{a}_t \leftarrow \pi_{\theta}(\cdot)$
- 11: Compute loss:  $\mathcal{L} = \text{Loss}(\hat{a}_t, a_t)$
- 12: Step policy parameters:  $\theta \leftarrow \theta - \nabla_{\theta} \mathcal{L}$
- 13: Decay dropout rate:  $p \leftarrow p_0 e^{-\lambda t}$
- 14: **end for**



Fig. 4: Random viewpoint dropout augmentation used during training. We randomly mask either the egocentric or third-person camera views with varying-sized noise patches.

### D. Deployment

During deployment, we study and support two configurations for handling the third-person input:

- 1) **With robot segmentation and inpainting**: Each overhead frame  $o_t^{\text{3rdR}}$  is first passed through the segmentation module  $o_t^{\text{3rd-mask}} = \text{SAM}_2(o_t^{\text{3rdR}})$  to identify the robot manipulator. This mask is then used by the inpainting module to blend in the static background reference:

$$o_t^{\text{3rd-masked}} = \text{Inpaint}(o_t^{\text{3rdR}}, o_{\text{ref}}^{\text{bg}}, o_t^{\text{3rd-mask}}). \quad (4)$$

This produces input observations that are visually consistent with the data distribution used during the training. Since SAM2 operates at around 40 FPS on

an A100, significantly faster than the action diffusion policy we train, the inpainting process incurs minimal computational overhead during deployment, and the real-time performance of the overall system is maintained.

- 2) **Without robot segmentation:** The raw third-person frame  $o_t^{3rdR}$  is directly provided to the policy without additional processing.

We compare the effectiveness of these two deployment modes in Section IV.

### E. Proposed Three-Jawed gripper



Fig. 6: Our custom three-jaw gripper system includes both a **Left:** handheld mode and a **Right:** motorized mode. Each jaw is about  $189 \text{ mm} \times 94 \text{ mm}$  in footprint size and is printed with variable infill such that most of the jaw is soft at 25% infill, except for the boxed areas. The gripping portion uses 5% infill to maximize grip and the portions attached to fixtures use a higher infill of 45% for stability.

We found the UMI gripper design to be challenging for tasks that required locking an object’s rotation. To address this, we designed a three-jaw gripper (Fig. 6) that uses two compliant jaws printed in TPU 95A and a rigid third jaw. Initially, all three jaws were designed to be flexible, but this required a much larger actuation force and resulted in unreliable performance. Adding a rigid jaw mitigated both issues. The usage of soft, compliant jaws removed the need for springs. For integration with a robotic manipulator, the gripping mechanism is actuated by a linear motor. The design

allows the gripper to carry payloads up to three times its own weight.

## IV. SYSTEM EVALUATION

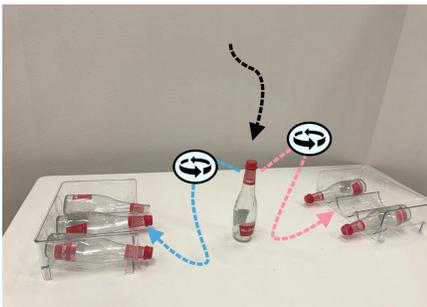
We evaluate the MV-UMI system across three key questions:

- 1) How effective is the third-person view integration for improving policy performance in tasks requiring context beyond the egocentric camera?
- 2) How important are the individual components of our data processing pipeline, such as viewpoint masking and inpainting, during training?
- 3) How necessary is inpainting at inference for maintaining performance during deployment, and can the policy operate effectively without it?

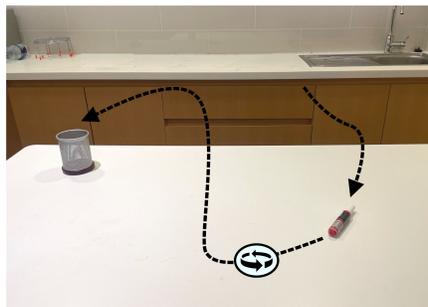
The tasks in Fig. 5 were selected to highlight scenarios where the egocentric view alone may be insufficient. Each task benefits from broader scene understanding through the third-person view and collectively involves challenges such as object rotations and payload handling. Specifically, the *Marker-Cup-Placer* task requires awareness of distant containers that are often outside the egocentric field of view at the start, as well as precise placement once the target cup leaves the frame.

The *Bottles-Rack-Inserter* task demands both dexterous alignment and payload handling capabilities that the UMI gripper design is unable to accomplish. One rack slot is left empty at random, requiring the model to locate it without egocentric visibility during the trajectory. Finally, the *Cans-Shelf-Placer* task emphasizes wide-scene context, as the shelf is moved to be outside the wrist camera’s view for most of the trajectory and on varying far positions on the table. In practice, we find this task challenging because the egocentric viewpoint goes out of distribution mid-execution in the changing room setting and is often occluded by the picked can.

1) *Experimental Setup:* A GoPro Hero-10 was used as the wrist-mounted gripper camera, and an Intel RealSense D455 was used to provide the third-person view. The third-person camera recorded continuously, while the episodes’ beginnings and ends were determined by the timestamps of



(a) *Bottles-Rack-Inserter:* Insert a glass bottle into the one rack with an empty slot.



(b) *Marker-Cup-Placer:* Pick and place a marker into a cup around the table.



(c) *Cans-Shelf-Placer:* Place a can on a shelf sitting on the other side of a table.

Fig. 5: Task descriptions in various environments for MV-UMI system.

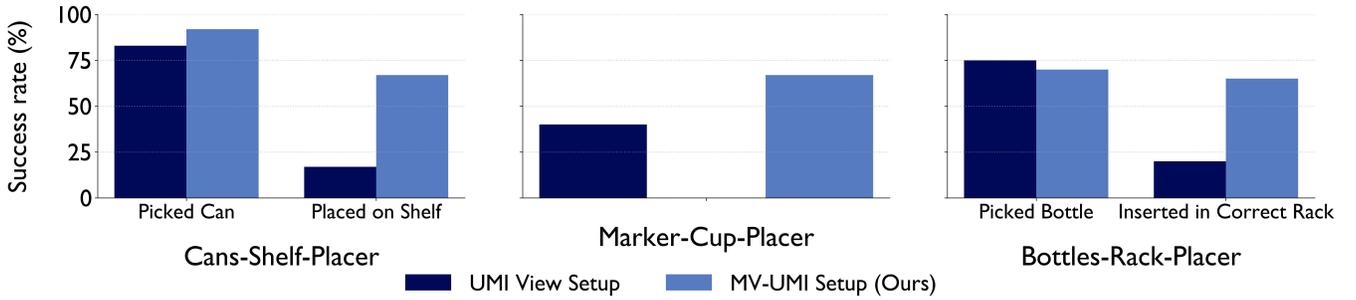


Fig. 7: Comparison of task success rates between *UMI View Setup* and our proposed *MV-UMI Setup*.

the GoPro camera videos. Scenes from our data-collection process can be seen in Fig. 3. The third-person footage was processed as detailed in Section III-D.

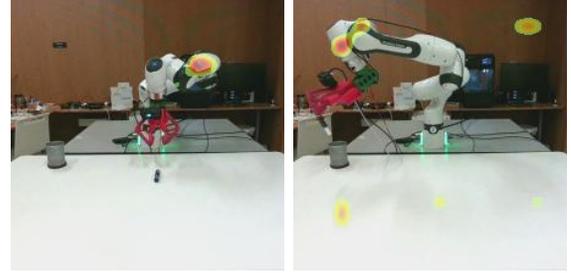
2) *Ablation Study*: In our ablation study, we found that the policy trained on human-aligned data failed to complete any task, despite often moving in sensible directions. We attribute this failure to the distribution shift in the observation space between training and deployment. By contrast, the policy trained on segmented data generalized well, even when the robot was not segmented at inference. This suggests that removing the human prevents the model from overfitting to spurious correlations between human motion and gripper actions, encouraging reliance on task-relevant scene features instead. Results from this study are summarized in Table II.

TABLE II: Ablation study (relative to MV-UMI success rate).

Ablation	Relative Performance
Without Human Segmentation	0.10
Without 3rd-Person View	0.60
Without Robot Inpainting	0.00
Without Robot Segmentation	0.80

3) *Attention Map Analysis*: To better understand the source of these differences, we analyzed attention maps from the Vision Transformer (ViT) encoder under both segmented and unsegmented training conditions. Using a forward hook on the multi-head self-attention layers, we captured the query-key attention scores for the class token, which indicate how strongly the global representation attends to individual image patches. These scores were reshaped into a spatial grid, normalized, and overlaid on the original frames to visualize regions of focus.

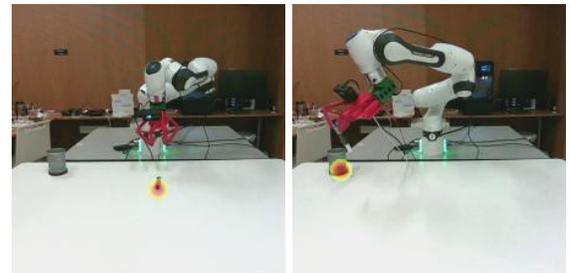
The visualizations (Fig. 8) directly support our ablation findings: models trained without human segmentation frequently direct attention toward embodiment-specific features such as the operator’s hand or robot arm, explaining their poor generalization. In contrast, our MV-UMI model consistently allocates attention to objects involved in the task, highlighting how data segmentation and inpainting guide the policy toward more object-centric manipulation. Interestingly, we also observe that the segmented model avoids attending to embodiment cues when the third-person view is left unsegmented during inference. We attribute this effect to the strong correlation between the human demonstrator’s motions and the corresponding robot actions; by removing



(a) Unsegmented model attending to embodiment and background features.



(b) MV-UMI model focusing on the object manipulated when the robot is segmented during runtime.



(c) MV-UMI model focusing on the object manipulated even when the robot is not segmented during runtime

Fig. 8: Attention maps from our ViT vision encoders under different training and inference conditions.

the human, the model is encouraged to focus on the true task-relevant signals.

4) *Task Results*: Fig. 7 shows the success rates of our multi-view setup compared to the single-camera UMI baseline. The egocentric wrist-mounted camera was sufficient in initial phases of tasks when the manipulated objects remained visible, but success rates drop sharply once the target leaves the field of view or occludes it after being picked. In contrast, our MV-UMI setup improves performance in these later stages.

For example, in the `Bottles-Rack-Insertor` task, the wrist-only setup typically succeeded in the picking phase but frequently failed during placement, as it lacked awareness of which side contained the empty slot. Similar, and in the `Marker-Cup-Placer` task, failures occurred when the cup moved outside the egocentric field of view, whereas MV-UMI maintained awareness of its position. Finally, in the `Cans-Shelf-Placer` task, the single-camera setup struggled when the shelf was positioned out of view for most of the trajectory.

5) *Limitations*: While the inclusion of a third-person view enhances scene understanding and improves task performance, it also introduces some challenges. First, occlusions caused either by the human demonstrator during training or by the robot during deployment can result in incomplete segmentation under the first deployment configuration described in Section III-D. However, we find that the second deployment configuration mitigates this issue during inference.

## V. CONCLUSION

In this paper, we introduced a data collection framework that addresses a significant limitation in current handheld demonstration systems: the restricted perspective of wrist-mounted cameras. Our multi-view approach integrates third-person camera footage with egocentric observations while mitigating domain shift through real-time person-masking and inpainting. This enables policies to gain broader contextual awareness, avoid spurious human-specific signals, and maintain cross-embodiment compatibility between demonstration and deployment.

## REFERENCES

- [1] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems*, Daegu, Republic of Korea, 2023.
- [2] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto, "Behavior generation with latent actions," in *International Conference on Machine Learning*, Vienna, Austria, 2024, pp. 26 991–27 008.
- [3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems*, Daegu, Republic of Korea, 2023.
- [4] N. M. Shafiullah, Z. Cui, A. A. Altanzaya, and L. Pinto, "Behavior transformers: Cloning  $k$  modes with one stone," in *Advances in neural information processing systems*, vol. 35, New Orleans, USA, 2022, pp. 22 955–22 968.
- [5] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Abu Dhabi, UAE, 2024, pp. 12 156–12 163.
- [6] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," in *Conference on Robot Learning*, Munich, Germany, 2024, pp. 1910–1924.
- [7] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," in *International Conference on Learning Representations*, Singapore, 2025.
- [8] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home," *arXiv preprint*, 2023.
- [9] 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," in *Robotics: Science and Systems*, Daegu, Republic of Korea, 2024.
- [10] M. Seo, H. A. Park, S. Yuan, Y. Zhu, and L. Sentis, "Legato: Cross-embodiment imitation using a grasping tool," *IEEE Robotics and Automation Letters*, vol. 10, no. 3, pp. 2854–2861, 2025.
- [11] Z. Wu, T. Wang, C. Guan, Z. Jia, S. Liang, H. Song, D. Qu, D. Wang, Z. Wang, N. Cao *et al.*, "Fast-UMI: A Scalable and Hardware-Independent Universal Manipulation Interface," *arXiv preprint*, 2024.
- [12] H. Etukuru, N. Naka, Z. Hu, S. Lee, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. M. Shafiullah, "Robot utility models: General policies for zero-shot deployment in new environments," in *IEEE International Conference on Robotics and Automation*, Atlanta, USA, 2025, pp. 8275–8283.
- [13] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Conference on Robot Learning*, Zurich, Switzerland, 2018, pp. 879–893.
- [14] S. Dass, W. Ai, Y. Jiang, S. Singh, J. Hu, R. Zhang, P. Stone, B. Abbatematteo, and R. Martín-Martín, "TeleMoMa: A Modular and Versatile Teleoperation System for Mobile Manipulation," in *RSS Workshop: Data Generation for Robotics*, Delft, Netherlands, 2024.
- [15] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," in *Robotics: Science and Systems*, Daegu, Republic of Korea, 2023.
- [16] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, "ACE: A Cross-platform and visual-Exoskeletons System for Low-Cost Dexterous Teleoperation," in *Conference on Robot Learning*, Munich, Germany, 2024, pp. 4895–4911.
- [17] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conference on Robot Learning*, Munich, Germany, 2024.
- [18] P. Praveena, G. Subramani, B. Mutlu, and M. Gleicher, "Characterizing input methods for human-to-robot demonstrations," in *ACM/IEEE International Conference on Human-Robot Interaction*, Daegu, Korea, 2019, pp. 344–353.
- [19] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," in *Robotics: Science and Systems*, Manchester, UK, 2021.
- [20] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the Wild: Learning 6 DoF Closed-Loop Grasping From Low-Cost Demonstrations," vol. 5, no. 3, 2020, pp. 4978–4985.
- [21] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "UMI-on-Legs: Making Manipulation Policies Mobile with Manipulation-Centric Whole-body Controllers," in *Conference on Robot Learning*, Munich, Germany, 2024, pp. 5254–5270.
- [22] 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," in *Robotics: Science and Systems*, Delft, Netherlands, 2024.
- [23] J. Duan, Y. R. Wang, M. Shridhar, D. Fox, and R. Krishna, "AR2-D2: Training a Robot Without a Robot," in *Conference on Robot Learning*, Atlanta, USA, 2023, pp. 2838–2848.
- [24] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "DROID: A large-scale in-the-wild robot manipulation dataset," in *Robotics: Science and Systems*, Delft, Netherlands, 2024.
- [25] M. Lepert, R. Doshi, and J. Bohg, "Shadow: Leveraging segmentation masks for cross-embodiment policy transfer," in *Conference on Robot Learning*, Munich, Germany, 2024.
- [26] L. Y. Chen, K. Hari, K. Dharmarajan, C. Xu, Q. Vuong, and K. Goldberg, "Mirage: Cross-embodiment zero-shot policy transfer with cross-painting," in *Robotics: Science and Systems*, Delft, Netherlands, 2024.
- [27] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," in *European Conference on Computer Vision*, Milan, Italy, 2024, pp. 306–324.
- [28] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," in *Conference on Robot Learning*, Munich, Germany, 2024, pp. 2475–2499.
- [29] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rüdle, C. Rolland, L. Gustafson *et al.*, "SAM 2: Segment Anything in Images and Videos," in *International Conference on Learning Representations*, 2024.
- [30] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial, and Multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.