Robotics: Science and Systems 2024
Delft, Netherlands, July 15-July 19, 2024

Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots

Cheng Chi*[1,2], Zhenjia Xu*[1,2], Chuer Pan[1], Eric Cousineau[3], Benjamin Burchfiel[3], Siyuan Feng[3], Russ Tedrake[3], Shuran Song[1,2]
[1]Stanford University, [2] Columbia University, [3]Toyota Research Insititute
https://umi-gripper.github.io

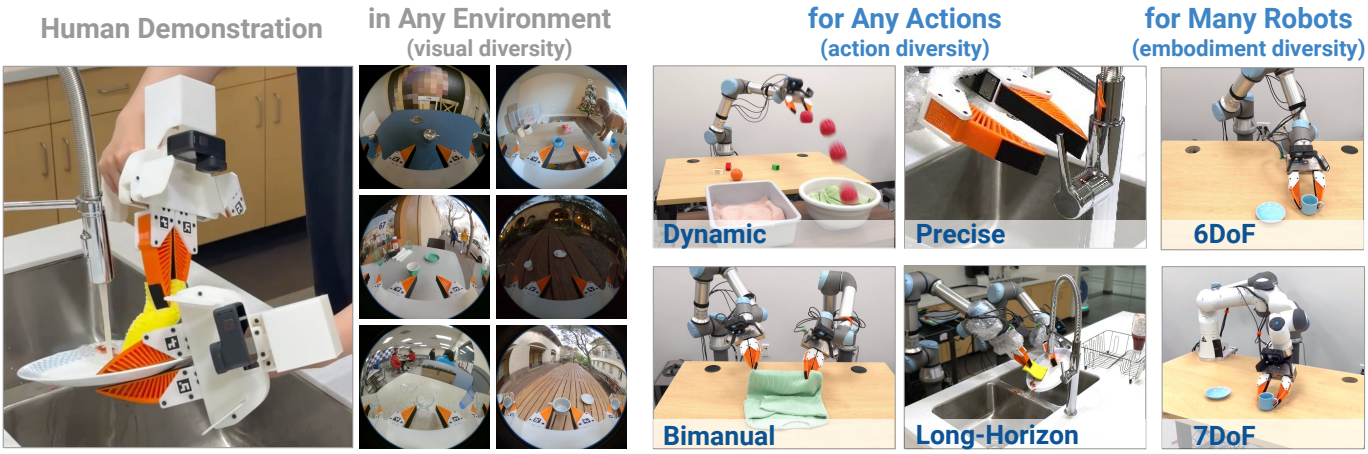


Fig. 1: **Universal Manipulation Interface (UMI)** is a portable, intuitive, low-cost data collection and policy learning framework. This framework allows us to transfer diverse human demonstrations to effective visuomotor policies. We showcase the framework for tasks that would be difficult with traditional teleoperation, such as dynamic, precise, bimanual and long-horizon tasks.

***Abstract*—We present Universal Manipulation Interface (UMI) – a data collection and policy learning framework that allows direct skill transfer from in-the-wild human demonstrations to deployable robot policies. UMI employs hand-held grippers coupled with careful interface design to enable portable, low-cost, and information-rich data collection for challenging bi-manual and dynamic manipulation demonstrations. To facilitate deployable policy learning, UMI incorporates a carefully designed policy interface with inference-time latency matching and a relative-trajectory action representation. The resulting learned policies are hardware-agnostic and deployable across multiple robot platforms. Equipped with these features, UMI framework unlocks new robot manipulation capabilities, allowing zero-shot generalizable dynamic, bimanual, precise, and long-horizon behaviors, by only changing the training data for each task. We demonstrate UMI's versatility and efficacy with comprehensive real-world experiments, where policies learned via UMI zero-shot generalize to novel environments and objects when trained on diverse human demonstrations. UMI's hardware and software system is open-sourced at https://umi-gripper.github.io.**

I. Introduction

How should we demonstrate complex manipulation skills for robots to learn from? Attempts in the field have approached this question primarily from two directions: collecting targeted in-the-lab robot datasets via teleoperation or leveraging unstructured in-the-wild human videos. Unfortunately, neither is sufficient, as teleoperation requires high setup costs for hardware and expert operators, while human videos exhibit a large embodiment gap to robots.

Recently, using sensorized hand-held grippers as a data collection interface [41, 51, 36] has emerged as a promising middle-ground alternative – simultaneously minimizing the embodiment gap while remaining intuitive and flexible. Despite their potential, these approaches still struggle to balance action diversity with transferability. While users can theoretically collect any actions with these hand-held devices, much of that data can not be transferred to an effective robot policy. As a result, despite achieving impressive **visual diversity** across hundreds of environments, the collected actions are constrained to simple grasping [41] or quasi-static pick-and-place [51, 36], lacking **action diversity**.

What prevents action transfer in previous work? We identified a few subtle yet critical issues:

- **Insufficient visual context:** While using a wrist-mounted camera is key for aligning the observation space and enhancing device portability, it restricts the scene's visual coverage. The camera's proximity to the manipulated object often results in heavy occlusions, providing insufficient visual context for action planning.
- **Action imprecision:** Most hand-held devices rely on monocular structure-from-motion (SfM) to recover robot actions. However, such methods often struggle to recover

∗ Indicates equal contribution

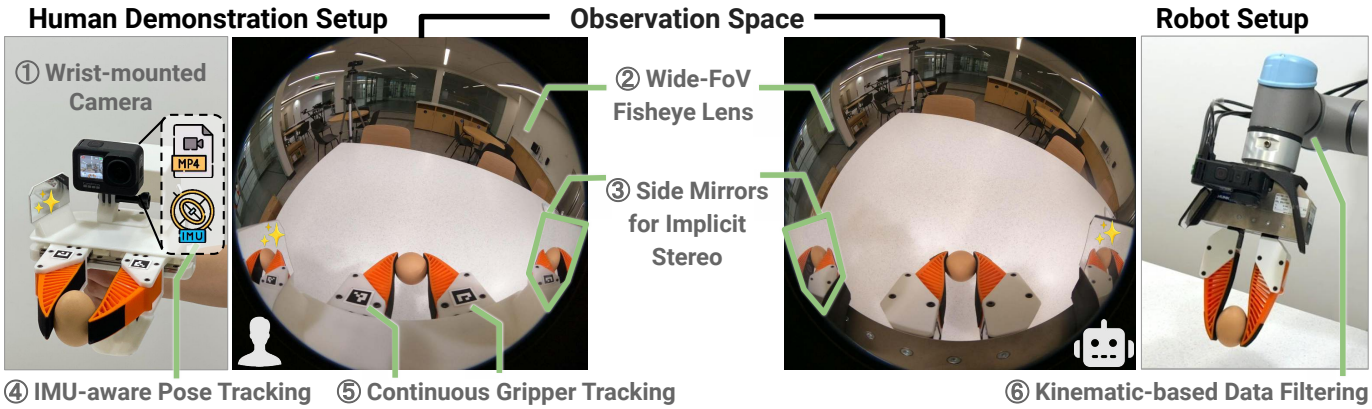


Fig. 2: **UMI Demonstration Interface Design.** Left: Hand-held grippers for data collection, with a GoPro as the only sensor and recording device. Middle: Image from the GoPro's 155° Fisheye view. Note the physical side mirrors highlighted in green which provide implicit stereo information. Right: UMI-compatible robot gripper and camera setup make observation similar to hand-held gripper view.

precise global action due to scale ambiguity, motion blur, or insufficient texture, which significantly restrict the precision of tasks for which the system can be employed.

- **Latency discrepancies:** During hand-held data collection, observation and action recording occur without latency. However, during inference, various latency sources, including sensor, inference, and execution latencies, arise within the system. Policies unaware of these latency discrepancies will encounter out-of-distribution input and in turn, generate out-of-sync actions. This issue is especially salient for fast and dynamic actions.
- **Insufficient policy representation:** Prior works often use simple policy representations (e.g., MLPs) with action regression loss, limiting their capacity to capture complex multimodal action distributions inherent in human data. Consequently, even with precisely recovered demonstrated actions and all discrepancies removed, the resulting policy could still struggle to fit the data accurately. This further hampers large-scale, distributed human data collection, as more demonstrators increase action multimodality.

In this paper, we address these issues with careful design of the demonstration and policy interface:

- First, we aim to identify the right **physical interface** for human demonstration that is intuitive and meanwhile able to capture all the information necessary for policy learning. Specifically, we use a Fisheye lens to increase the field of view and visual context, and add side mirrors on the gripper to provide implicit stereo observation. When combined with the GoPro's built-in IMU sensor, we can enable robust tracking under fast motion.
- Second, we explore the right **policy interface** (i.e., observation and action representations) that could make the policy hardware-agnostic and thereby enable effective skill transfer. Concretely, we employ inference-time latency matching to handle different sensor observation and execution latency, use relative trajectory as action representation to remove the need for precise global action, and finally, apply Diffusion Policy [9] to model multimodal action distributions.

The final system, **Universal Manipulation Interface (UMI)**, provides a practical and accessible framework to unlock new robot manipulation skills, allowing us to demonstrate any actions in any environment while maintaining high transferability from human demonstration to robot policy.

With just a wrist-mounted camera on the hand-held gripper (Fig. 2), we show that UMI is capable of achieving a wide range of manipulation tasks that involve dynamic, bimanual, precise and long-horizon actions by only changing the training data for each task (Fig. 1). Furthermore, when trained with diverse human demonstrations, the final policy exhibits zero-shot generalization to novel environments and objects, achieving a remarkable 70% success rate in out-of-distribution tests, a level of generalizabilty seldomly observed in other behavior cloning frameworks. We open-source the hardware and software system at https://umi-gripper.github.io.

II. Related Works

A key enabler for any data-driven robotics system is the data itself. Here, we review a few typical data collection workflows in the context of robotic manipulation.

A. Teleoperated Robot Data

Imitation learning learns policies from expert demonstrations. Behavior cloning (BC), utilizing teleoperated robot demonstrations, stands out for its direct transferability. However, teleoperating real robots for data collection poses significant challenges. Previous approaches utilized interfaces such as 3D spacemouse [9, 55], VR or AR controllers [35, 3, 13, 19, 31, 52, 12], smartphones [44, 46, 22], and haptic devic [38, 48, 43, 26, 4] for teleoperation. These methods are either very expensive or hard to use due to high latency and lack of user intuitiveness. While recent advancements in leader-follower (i.e. puppetting) devices such as ALOHA [54, 15] and GELLO [47] offer promise with intuitive and low-cost interfaces, their reliance on real robots during data collection limits the type and number of environments the system can gain access to for "in-the-wild" data acquisition. Exoskeletons [14, 20] remove the dependence on real robots during data

collection, however, require fine-tuning using teleoperated real robot data for deployment. Moreover, the resulting data and policy from aforementioned devices are embodiment-specific, preventing reusage for different robots.

In contrast, UMI eliminates the need for physical robots during data collection and offers a more portable interface for in-the-wild robot teaching, providing data and policies that are transferable to different robot embodiments (e.g., 6DoF or 7DoF robot arms).

B. Visual Demonstrations from Human Video

There's a distinct line of work dedicated to policy learning from in-the-wild video data (e.g. YouTube videos). The most common way is to learn from diverse passive human demonstration videos. Utilizing passive human demonstrations, previous works learn task cost functions [37, 8, 1, 21], affordance functions [2], dense object descriptors [40, 24, 39], action correspondences [33, 28], and pre-trained visual representations [23, 49].

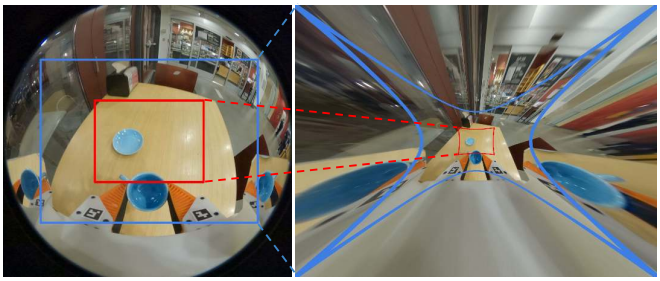
However, this approach encounters three major challenges. Firstly, most video demonstrations lack explicit action information, crucial for learning generalizable policies. To infer action data from passive human video, previous works resort to hand pose detectors [44, 1, 38, 28], or combining human videos with in-domain teleoperated robot data to predict actions [33, 20, 34, 28]. Second, the evident embodiment gap between humans and robots hinders action transfer. Efforts to bridge the gap include learning human-to-robot action mapping with hand pose retargetting [38, 28, 45] or extracting embodiment-agnostic keypoints [50]. Despite these attempts, the inherent embodiment differences still complicate policy transfer from human video to physical robots. Thirdly, the inherent observation gap induced by the embodiment gap in this line of work introduces inevitable mismatch between train/inference time observation data, exacerbating the transferability of the resulting policies, despite efforts in aligning demonstration observation with robot observation [20, 28].

In contrast, data collected with UMI exhibit minimal embodiment gap both in action and observation spaces, enabled by precise manipulation action extraction via robust visual-inertial camera tracking and the shared Fisheye wrist-mounted cameras during teaching and testing. Consequently, this enables in-the-wild zero-shot policy transfer for dynamic, bimanual, precise, and long-horizon manipulation tasks.

C. Hand-Held Grippers for Quasi-static Actions

Hand-held grippers [41, 51, 10, 32, 27, 25] minimize observation embodiment gaps in manipulation data collection, offering portability and intuitive interfaces for efficient data collection in the wild. However, accurately and robustly extracting 6DoF end-effector (EE) pose from these devices remains challenging, hindering the deployment of robot policies learned from these data on fine-grained manipulation tasks.

Prior works attempted to address this issue through various approaches, such as SfM [51, 25] which suffers from scale ambiguity; RGB-D fusion [41] which requires expensive sensors and onboard compute; external motion tracking [32, 27] which is limited to lab settings. These devices, constrained to quasi-static actions due to low EE tracking accuracy and robustness, often necessitate cumbersome onboard computer or external motion capture (MoCap) systems, diminishing their feasibility for in-the-wild data collection. In contrast, UMI integrates state-of-the-art SLAM [6] with built-in IMU data from GoPro, to accurately capture 6DoF actions at the global scale. The high-accuracy data enables trained BC policy to learn bimanual tasks. With thorough latency matching, UMI further enables real-world deployable policy for dynamic actions such as tossing.



(a) Raw Fisheye Image (b) Rectified Image

Fig. 3: **Fisheye vs Rectilinear** (a) UMI policies use raw Fisheye image as observation. (b) Rectifying a large 155° FoV image to the pin-hole model severely stretches the peripheral view (outside of blue line), while compresses the most important information at the center to a small area (inside of red line).

Recently, Dobb-E [36] proposed a "reacher-grabber" tool mounted with an iPhone to collect single-arm demonstrations for the Stretch robot. Yet, Dobb-E only demonstrates policy deployment for quasi-static tasks and requires environment-specific policy fine-tuning. Conversely, using only data collected with UMI enables trained policy to zero-shot generalize to novel in-the-wild environments, unseen objects, multiple robot embodiments, for dynamic, bimanual, precise and long-horizon tasks.

III. Method

Universal Manipulation Interface (UMI) is hand-held data collection and policy learning framework that allows direct transfer from in-the-wild human demonstrations to deployable robot policies. It is designed with the following goals in mind:

- **Portable.** The hand-held UMI grippers can be taken to any environment and start data collection with close-to-zero setup time.
- **Capable.** The ability to capture and transfer natural and complex human manipulation skills beyond pick-and-place.
- **Sufficient.** The collected data should contain sufficient information for learning effective robot policies and contain minimal embodiment-specific information that would prevent transfer.
- **Reproducible**: Researchers and enthusiasts should be able to consistently build UMI grippers and use data to train their own robots, even with different robot arms.

The following sections describe how we enable the above goals through our hardware and policy interface design.

A. Demonstration Interface Design

UMI's data collection hardware takes the form of a trigger-activated, handheld 3D printed parallel jaw gripper with soft fingers, mounted with a GoPro camera as the **only** sensor and recording device (see HD1). For bimanual manipulation, UMI can be trivially extended with another gripper. The key research question we need to address here is:

How can we capture sufficient information for a wide variety of tasks with just a wrist-mounted camera?

Specifically, on the **observation** side, the device needs to capture sufficient visual context to infer action HD2 and critical depth information HD3. On the **action** side, it needs to capture precise robot action under fast human motion HD4, detailed subtle adjustments on griping width HD5, and automatically check whether each demonstration is valid given the robot hardware kinematics HD6. The following sections describe details on how we achieve these goals.

HD1. Wrist-mounted cameras as input observation. We rely solely on wrist-mounted cameras, without the need for any external camera setups. When deploying UMI on a robot, we place GoPro cameras with the same location with respect to the same 3D-printed fingers as on the hand-held gripper. This design provides the following benefits:

1) **Minimizing the observation embodiment gaps.** Thanks to our hardware design, the videos observed in wrist-mount cameras are almost indistinguishable between human demonstrations and robot deployment, making the policy input less sensitive to embodiment.
2) **Mechanical robustness.** Because the camera is mechanically fixed relative to the fingers, mounting UMI on robots does not require camera-robot-world calibration. Hence, the system is much more robust to mechanical shocks, making it easy to deploy.
3) **Portable hardware setup.** Without the need for an external static camera or additional onboard compute, we largely simplify the data collection setup and make the whole system highly portable.
4) **Camera motion for natural data diversification.** A side benefit we observed from experiments is that when training with a moving camera, the policy learns to focus on task-relevant objects or regions instead of background structures (similar in effect to random cropping). As a result, the final policy naturally becomes more robust against distractors at inference time.

Avoiding use of external static cameras also introduce additional challenges for downstream policy learning. For example, the policy now needs to handle non-stationary and partial observations. We mitigated these issues by leveraging wide-FoV Fisheye Lens HD2, and robust visual tracking HD4, described in the following sections.

HD2. Fisheye Lens for visual context. We use a 155-degree Fisheye lens attachment on wrist-mounted GoPro camera, which provides sufficient visual context for a wide range of tasks, as shown in Fig. 2. As the policy input, we directly use raw Fisheye images *without undistortion* since Fisheye effects conveniently preserve resolution in the center while compressing information in the peripheral view. In contrast, rectified pinhole image (Fig. 3 right) exhibits extreme distortions, making it unsuitable for learning due to the wide FoV. Beyond improving SLAM robustness with increased visual features and overlap [53], our quantitative evaluation (Sec V-A) shows that the Fisheye lens improves policy performance by providing the necessary visual context.

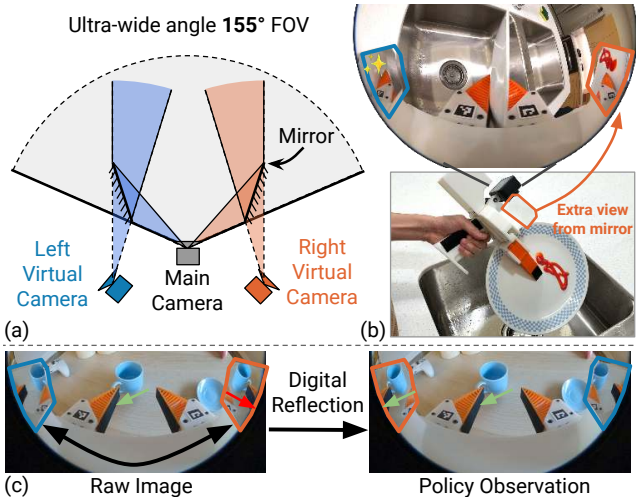


Fig. 4: **UMI Side Mirrors.** The ultra-wide-angle camera coupled with strategically positioned mirrors, facilitates implicit stereo depth estimation. **(a)**: The view through each mirror effectively creates two virtual cameras, whose poses are reflected along the mirror planes with respect to the main camera. **(b)**: Ketchup on the plate, occluded from the main camera view, is visible inside the right mirror, proving that mirrors simulate cameras with different optical centers. **(c)**: We digitally reflect the content inside mirrors for policy observation. Note the orientation of the cup handle becomes consistent across all 3 views after reflection.

HD3. Side mirrors for implicit stereo. To mitigate the lack of direct depth perception from the monocular camera view, we placed a pair of physical mirrors in the cameras' peripheral view which creates implicit stereo views all in the same image. As illustrated in Fig 4 (a), the images inside the mirrors are equivalent to what can be seen from additional cameras reflected along the mirror plane, without the additional cost and weight. To make use of these mirror views, we found that digitally reflecting the crop of the images in the mirrors, shown in Fig 4 (c), yields the best result for policy learning (Sec. V-A). Note that without digital reflection, the orientation of objects seen through side mirrors is the opposite of that in the main camera view.

HD4. IMU-aware tracking. UMI captures rapid movements with absolute scale by leveraging GoPro's built-in capability to record IMU data (accelerometer and gyroscope) into standard mp4 video files [18]. By jointly optimizing visual tracking and inertial pose constraints, our Inertial-monocular SLAM system based on ORB-SLAM3 [7] maintains tracking for a short period of time even if visual tracking fails due to motion blur or a lack of visual features (e.g. looking down at a table). This allows UMI to capture and deploy highly

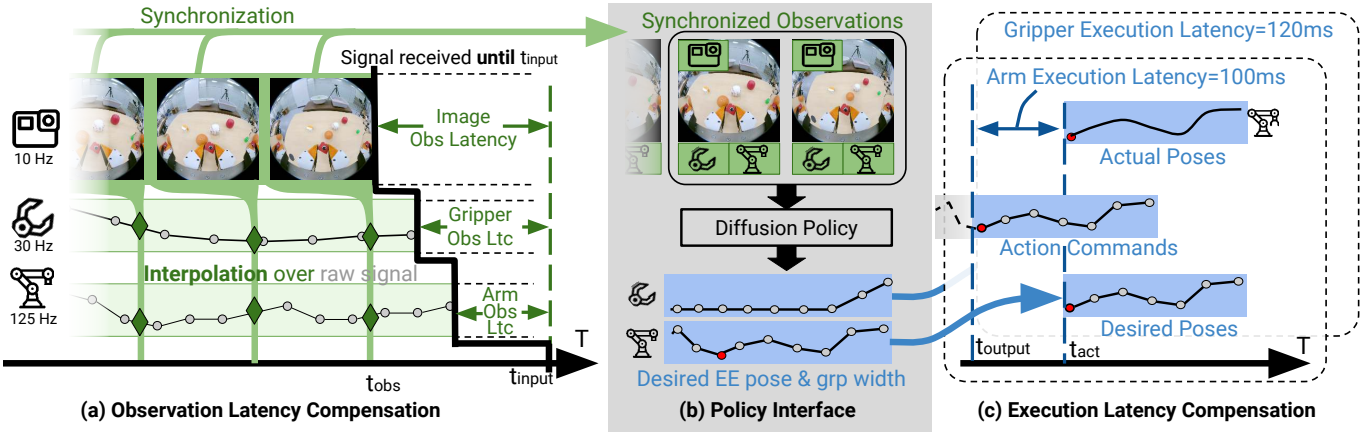


Fig. 5: **UMI Policy Interface Design.** (b) UMI policy takes in a sequence of **synchronized** observations (RGB image, relative EE pose, and gripper width) and outputs a sequence of **desired** relative EE pose and gripper width as action. (a) We synchronize different observation streams with physically measured latencies. (c) We send action commands ahead of time to compensate for robots' execution latency.

dynamic actions such as tossing (shown in Fig 7). In addition, the joint visual-inertial optimization allows direct recovery of real metric scale, important for action precision and inter-gripper pose proprioception PD2.3: a critical ingredient to enable bimanual policy.

HD5. Continuous gripper control. In contrast to the binary open-close action used in prior works [41, 44, 55], we found commanding gripper width continuously significantly expands the range of tasks doable by parallel-jaw grippers. For example, the tossing task (Fig. 7) requires precise timing for releasing objects. Since objects have different widths, binary gripper actions will be unlikely to meet the precision requirement. On UMI gripper, finger width is continuously tracked via fiducial markers [16] (Fig. 2 left). Using series-elastic end effectors principle [42], UMI can implicitly record and control grasp forces by regulating the deformation of soft fingers through continuous gripper width control.

HD6. Kinematic-based data filtering. While the data collection process is robot-agnostic, we apply simple kinematic-based data filtering to select valid trajectories for different robot embodiments. Concretely, when the robot's base location and kinematics are known, the absolute end-effector pose recovered by SLAM allows kinematics and dynamics feasibility filtering on the demonstration data. Training on the filtered dataset ensures policies comply with embodiment-specific kinematic constraints.

Putting everything together. The UMI gripper weighs 780g, with an external dimension of $L310mm \times W175mm \times H210mm$ and finger stroke of $80mm$. The 3D printed gripper has a BoM cost of **$73**, while the GoPro camera and accessories total **$298**. As shown in Fig. 2, we can equip any robot arms with a compatible gripper and camera setup.

B. Policy Interface Design

With the collected demonstration data, we can train a visuomotor policy that takes in a sequence of synchronized observations (RGB images, 6 degrees-of-freedom end-effector pose, and gripper width) and produces a sequence of actions (end-effector pose and gripper width) as shown in Fig. 5 (b). In this paper, we use Diffusion Policy [9] for all of our experiments, while other frameworks such as ACT [54] could potentially serve as a drop-in replacement.

An important goal of UMI's policy interface design is to ensure the interface is **agnostic to underlying robotic hardware platforms** such that the resulting policy, trained on one data source (i.e., hand-held gripper), could be directly deployed to different robot platforms. To do so, we aim to address the following two key challenges:

- **Hardware-specific latency.** The latency of various hardware (streaming camera, robot controller, industrial gripper) is highly variable across system deployments, ranging from single-digit to hundreds of milliseconds. In contrast, all information streams captured by UMI grippers have zero latency with respect to the image observation, thanks to GoPro's synchronized video, IMU measurements and the vision-based gripper width estimation.
- **Embodiment-specific proprioception.** Commonly used proprioception observations such as joint angles and EE pose are only well-defined with respect to a specific robot arm and robot base placement. In contrast, UMI needs to collect data across diverse environments and be generalizable to multiple robot embodiments.

In the following sections, we will describe three policy interface designs that address these challenges.

PD1. Inference-time latency matching. While UMI's policy interface assumes synchronized observation streams and immediate action execution, physical robot systems do not conform to this assumption. If not carefully handled, the timing mismatch between training and testing can cause large performance drops on dynamic manipulation tasks that require rapid movement and precise hand-eye coordination, demonstrated in Sec V-B. In this paper, we separately handle timing discrepancies on the observation and action sides:

PD1.1) Observation latency matching. On real robotic systems, different observation streams (RGB image, EE pose, gripper width) are captured by distributed micro-controllers, resulting in different observation latency.

For each observation stream, we individually measure their latency (details see §A1-A3). At inference time, we align all observations with respect to the stream with the highest latency (usually the camera). Specifically, we first temporally downsample the RGB camera observations to the desired frequency (often 10-20Hz), and then use the capture timestamp of each image t_{obs} to linearly interpolate gripper and robot proprioception streams. In bimanual systems, we soft-synchronize two cameras by finding the nearest neighbor frames, which can be off by a maximum of $\frac{1}{60}$ seconds. The result is a sequence of synchronized observations that conform to UMI policy, shown in Fig. 5 (a).

PD1.2) Action latency matching. UMI policy assumes the output as a sequence of synchronized EE poses and gripper widths. However, in practice, robot arms and grippers can only track the desired pose sequence up to an **execution latency**, that varies across different robot hardware. To make sure the robots and grippers reach the desired pose at the desired time (given by the policy), we need to send commands **ahead of time** to compensate for execution latency, as shown in Fig. 5 (c). See §A4 for execution latency calibration details.

During execution, the UMI policy predicts the action sequence starting at the last step of observation t_{obs}. The first few actions predicted are immediately outdated due to observation latency $t_{input} - t_{obs}$, policy inference latency $t_{output} - t_{input}$ and execution latency $t_{act} - t_{output}$. We simply discard the outdated actions and only execute actions with the desired timestamp after t_{act} for each hardware.

PD2. Relative end-effector pose. End-effector (EE) pose is central to both UMI's observation and action space. To avoid dependence on embodiment/deployment-specific coordinates, we represent all EE poses relative to gripper's current EE pose.

PD2.1) Relative EE trajectory as action representation. Prior works have shown the significant impact of action space selection on task performance [9], with experimental evidence favoring absolute positional actions over delta actions. However, we found that a relative trajectory representation, defined for an action sequence starting at t_0 as a sequence of $SE(3)$ transforms denoting the desired pose at t relative to the initial EE pose at t_0, allows the system to be more robust against tracking errors during data collection and camera displacements.

PD2.2) Relative EE trajectory as proprioception. Similarly, we represent the proprioception of history EE poses as a relative trajectory. When observation horizon is set to 2, this representation effectively provides velocity information to the policy. Combined with our wrist-mounted camera observation space, relative trajectory allows our system to be **calibration-free**. Moving the robot base during execution will not affect task performance (Fig. 10 (a)), as long as the objects are still within reach range, making the UMI framework applicable to mobile manipulators as well.

PD2.3) Relative inter-gripper proprioception. When using UMI in a bimanual setup, we found that providing the policy with the relative pose between the two grippers to be critical for bimanual coordination and task success, as shown in Sec. V-C. The effect of inter-gripper proprioception is particularly large when the visual overlap between two cameras is small. The inter-gripper proprioception is enabled by our map-then-localize data collection scheme that constructs a scene-level global coordinate system HD4. For each new scene, we first collect a video that builds a map for the scene. Then, all demonstrations collected in this scene are relocalized to the same map, therefore sharing the same coordinate system. Despite the videos from each gripper being relocalized separately, the relative pose between two grippers at each time step can be calculated using their shared coordinates.

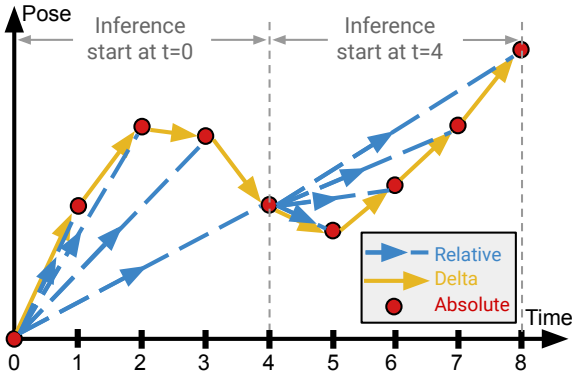


Fig. 6: **Relative Trajectory as Action Representation**. Relative trajectory, used by UMI, is a sequence of end-effector (EE) poses relative to **the same current EE pose** for each inference step. In contrast, Delta action represents each action step relative to its immediate previous action, therefore accumulates error. Absolute action requires a global coordinate frame for all actions, which is difficult to define for in-the-wild data collection.

IV. Evaluations

In our experiment, we aim to evaluate the UMI framework's effectiveness for deployable policy learning in three aspects:

- **Capability:** How well can we transfer UMI demonstrations to effective robot policy? Especially for complex, dynamic, bimanual, and long-horizon manipulation skills.
- **Generalization:** Will data collected in the wild within diverse environments help the policy to generalize to unseen environments and objects?
- **Data collection efficiency:** How fast can we collect manipulation data with UMI? What's the accuracy of the SLAM system?

To access capability and generalization, we evaluate UMI on 4 real-world robotic tasks across both narrow domain and in-the-wild environments, shown in Fig. 7. To measure data collection efficiency, we compare the UMI gripper with human hand demonstration and a typical teleop interface. See §B for detailed data collection protocol.

V. Capability Experiments

We study UMI's ability to capture and transfer single-hand, bimanual, dynamic, and long-horizon manipulation skills with four tasks. For capability experiments, all tasks are evaluated in the same environment as data collection but with randomized robot and object initial states. To ensure a fair comparison,

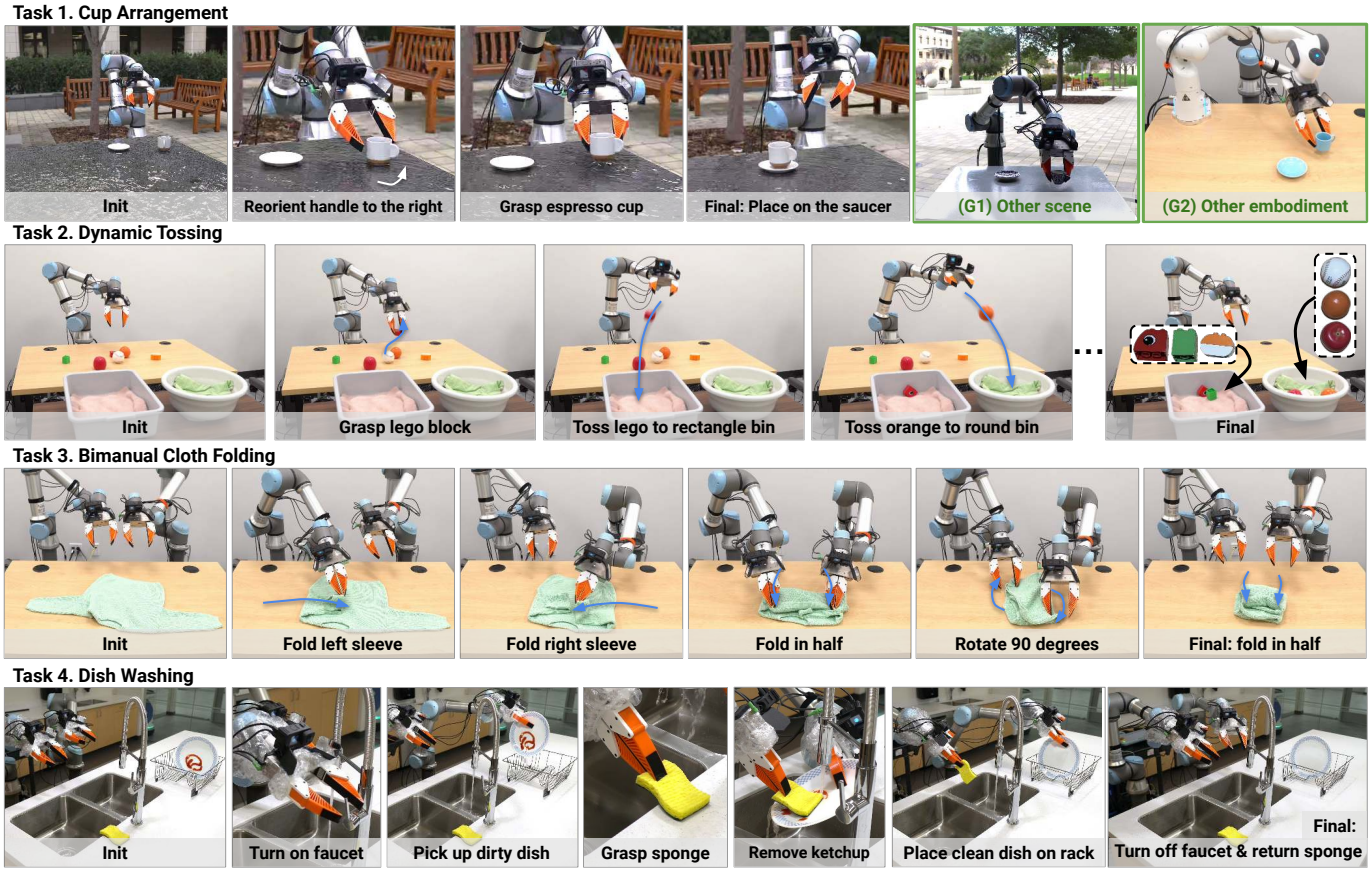


Fig. 7: **Policy Rollouts.** We test UMI on a variety of challenging real-world tasks. **Cup arrangement** tests UMI's ability to learn both prehensile and non-prehensile actions, and to capture multi-modal action distribution (clockwise and counter-clockwise rotation). This task is evaluated in both narrow-domain and unseen environments as well as two robot embodiments. **Dynamic tossing** tests UMI's ability to capture and transfer rapid human motions as well as precise hand-eye coordination. **Bimanual cloth folding** tests UMI's ability to synchronize two-arm coordination. **Dish washing** tests UMI's ability to handle long-horizon tasks that involve multiple rigid, deformable, and articulated objects. Please check videos on our website for more details.

we use exactly *the same initial state across all methods* for both the robot and objects, by manually aligning the scene against pre-recorded images. See §C for detailed evaluation protocol and videos for all experiments.

A. *Cup Arrangement*

Task Place an espresso cup on the saucer with its handle facing to the left of the robot, Fig. 7. We defined task success as when the cup is placed upright on the saucer with its handle within $\pm 15°$ to the left.

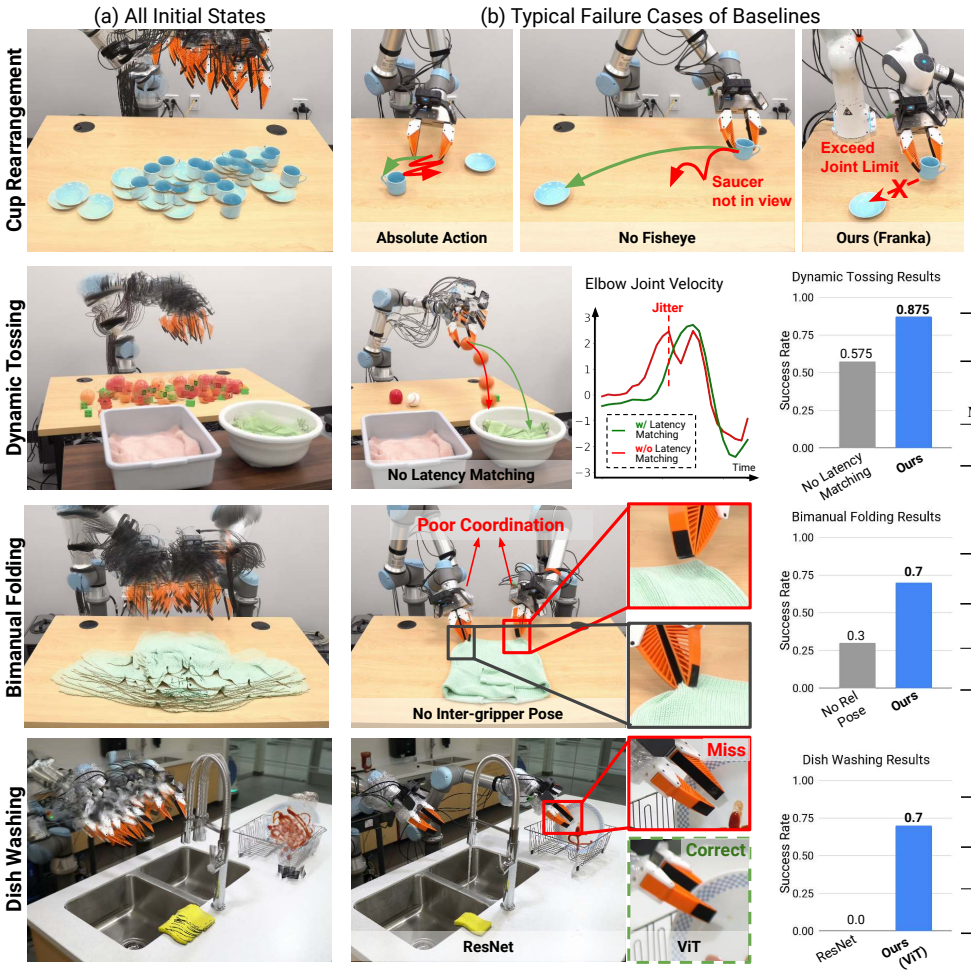
Capability (what makes the task difficult?) This task tests the system's ability to learn both prehensile (pick and place) and non-prehensile actions (i.e., pushing to reorientate the cup). When the handle faces straight away from the robot, the two equally valid solutions: rotation clockwise and counter-clockwise form a multi-modal action distribution. This task also tests UMI's ability to sense relative depth through monocular camera observation and side mirrors.

Performance The training dataset contains 305 episodes collected by 2 demonstrators, evaluation includes 20 test cases, with the testing initial state distribution shown in Fig. 8 (a). UMI can complete the task 20/20. The next paragraphs will discuss our ablation studies around our key design decisions.

Cross-robot generalization: To demonstrate UMI's cross-embodiment generality, we also deployed the same policy checkpoint on a Franka Emika FR2 robot, shown in Fig. 1 and Fig. 8. This experiment achieves $18/20 = 90\%$ success rate, with the 2 failure cases being joint limit violations, which could have been avoided if we had mounted the FR2 robot at a different location.

No Fisheye lens [HD2]: To ablate the importance of having a wide field-of-view (FoV) Fisheye lens, we post-processed the dataset by rectifying and cropping each image to a square with 69° horizontal and vertical FoV. This is a generous analogy of RealSense D415 (69° HFoV, 42° VFoV) and iPhone wide camera (69° HFoV, 51° VFoV). This baseline only achieves $11/20 = 55\%$ success rate. Beyond the expected failure mode where the cup is outside of camera view, we found this baseline policy to perform surprisingly poor even if the object is visible, with often jittery motions. We suspect that during training, the poor object visibility forced the policy to be unnecessarily multimodal.

Alternative action spaces [PD2]: As alternatives to our relative trajectory as action representation, we also consider absolute and delta action spaces as illustrated in Fig 6. Since



Per-object Success Rate

	Apple	Baseball	Orange	Green Block	Red Block	Orange Block	Overall
No Latency Matching	0.70	0.50	0.70	0.30	0.65	0.60	0.575
Ours	**0.80**	**0.85**	**0.85**	**1.00**	**0.90**	**0.85**	**0.875**

Substep Success Rate

	Right Sleeve	Left Sleeve	Fold Bottom	Rotate	Fold Final	Overall
No Rel Pose	**0.90**	0.70	0.35	0.40	0.30	0.30
Ours	**0.90**	**0.90**	**0.90**	**1.00**	**0.80**	**0.70**

Substep Success Rate

	Turn Faucet	Grasp Dish	Grasp Sponge	Wash Dish	Place Dish	Overall
ResNet	0.50	0.20	0.90	0.00	0.00	0.00
Ours (ViT)	**1.00**	**0.90**	**0.95**	**0.75**	**0.75**	**0.70**

Fig. 8: **Narrow-domain Evaluation Results.** (a) Initial states for all evaluation episodes overlayed together. For each task, all methods start with the same set of initial states, matched manually with reference images. (b) Typical failure mode of the baseline/ablation policy. The red arrow indicates failure behavior, green arrow indicates desired behavior. (c) Success rate over 20 evaluation episodes, best performance for each column are bolded. Please check our website for more comparison videos.

the SLAM system outputs pose relative to the first frame of the mapping video (details in §D), we can only calculate relative and delta actions directly using SLAM output. To compute absolute actions in the robot base frame, we calibrate both SLAM coordinates and the robot with respect to the same fiducial markers [16] placed on the table.

The delta action baseline achieves $16/20 = 80\%$ success rate. The absolute action baseline performs surprisingly poorly with only $5/20 = 25\%$ success rate, demonstrating a noticeable bias in action selection, likely due to inaccurate calibration between the SLAM and robot base coordinate frames (Fig. 8 (b)). While theoretically the performance of this baseline could approach that of relative trajectory with better calibration, this experiment underscores the difficulty of obtaining action data with absolute coordinates, even in controlled lab settings.

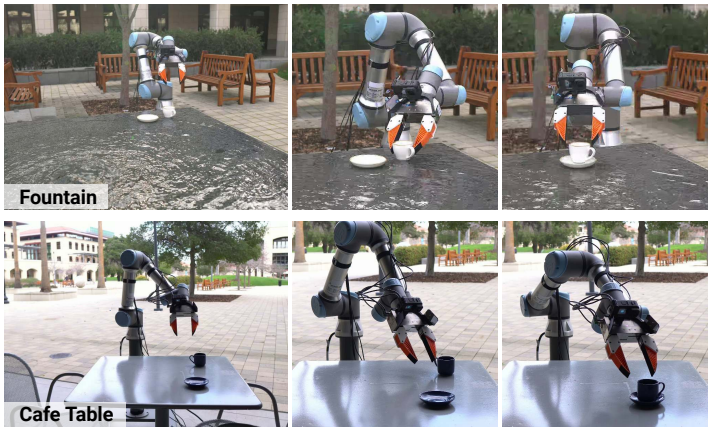
Effect of side mirrors [HD3]: To our surprise, directly providing mirror images decreases the performance from $18/20 = 90\%$ (no mirror) to $17/20 = 85\%$. To fully take advantage of side mirrors, we need to digitally reflect the content inside mirrors and swap left and right mirror images, which achieves a $20/20 = 100\%$ success rate. We hypothesize that without digital reflection, the opposite motions observed in the main and mirrored images might confuse vision encoders, especially those with translational equivariance.

B. *Dynamic Tossing*

Task The robot is tasked to sort 6 objects from the YCB object set [5] randomly placed on a table by tossing them to the corresponding bin. The 3 spherical objects (baseball, orange, apple) should be tossed into the round bin, while the 3 Lego Duplo pieces go into the rectangular bin (Fig. 7). The bins are placed beyond the robot's kinematic reach range to highlight the necessity of dynamic action for this task.

Capability: The dynamic tossing task demonstrates UMI's ability to capture and transfer fluid and rapid human motions, precise hand-eye coordination (between RGB and proprioception) and timing alignment (between robot and gripper).

Performance: We collected 280 demonstration episodes for this task, with mixed multi and single-object picking and tossing. Our policy (with inference time latency matching) achieves $105/120 = 87.5\%$ success rate, counted by the number of objects successfully tossed to their corresponding bin.



(a) Generalization to **novel environments** and **novel cups**

(b) Typical Failure Case of Policy using **Narrow-domain** data

		Cafe Table	Water Fountain	Success Rate
CLIP ViT finetune with narrow-domain Data		0 / 10	0 / 10	0.0
In-the-wild Data	Training Cup	16 / 25	12 / 15	**0.7**
	Unseen Cup	6 / 10	9 / 10	**0.75**
	Overall	22 / 35	21 / 25	**0.72**

(c) Quantitative Comparison in Unseen Environments

Fig. 9: **In-the-wild Cup Arrangement Evaluation.** (a) The policy, trained with in-the-wild data, demonstrates strong generalization capability to novel environments and novel objects. (b) With only narrow-domain data, the policy struggles to generalize to new environments. (c) Comprehensive qualitative results are provided, and additional comparison videos can be found on our website.

No Latency Matching [PD1]: With the same trained policy, we disable inference-time latency matching by setting the measured latencies for all observation and action streams to 0. We visually observe the policy's movement is much more jittery due to the out-of-sync observations and executions.

While the jitteriness minimally affects grasping, its impact on tossing performance is notable as it disrupts the robot motion to achieve the desired tossing velocity, as illustrated in the elbow joint velocity curve in Fig. 8. In addition, the misalignment between the gripper and robot action (due to their different execution latency) leads to suboptimal object release during tossing. As a result, the final success rate decreased to $69/120 = 57.5\%$.

C. *Bimanual Cloth Folding*

Task Two robot arms need to coordinate and fold the sweater's sleeves inward, fold up the bottom hem, rotate 90 degrees, and finally fold the sweater in half again (Fig. 7). See §C for detailed evaluation protocol.

Capability Manipulating high degrees of freedom deformable objects is challenging due to their complex dynamics and underactuation. In addition, this task requires tight coordination between arms. For example, lifting the bottom hem requires two arms to pick it up at the same time, and if one of the arm is just slightly too fast or slow this step will fail.

Performance We collected 250 demonstrations from two demonstrators, with randomized initial states by translating/rotating the shirt and folding the sleeves. We use a single centralized policy to generate actions for both robot arms and grippers. Our policy achieves $14/20 = 70\%$ success rate.

No relative inter-gripper proprioception [PD2.3]: Without inter-gripper proprioception information (during both training and eval), the coordination between the two arms becomes significantly worse. The most salient failure case is when the two arms lift the bottom hem of the shirt, where the baseline policy often misses one of the grasps due to asynchronous grasp action (Fig. 8 (b)). As a result, the baseline policy only achieves success rate of $6/20 = 30\%$. In contrast, UMI policy synchronizes the grasp by first reaching the pre-grasp pose and waiting until both arms are in position before simultaneously grasping and folding.

D. *Dish Washing*

Task The robot needs to execute **7 steps** of sequentially dependent actions (turn on faucet, grasp plate, pick up sponge, wash and wipe plate until ketchups are removed, place plate, place the sponge and turn off faucet), shown in Fig. 7. See §C for detailed evaluation protocol.

Capability This task pushes the boundaries of robot manipulation capability from several fronts: 1) it is an **ultra-long horizon** task where each step's success depends on the previous one; 2) the robot needs to perceive and manipulate **complex fluid** including both Newtonian fluid (i.e., water) and non-Newtonian fluid (i.e., ketchup). 3) the wiping motion requires using a **deformable tool** (i.e., sponge) while coordinating both arms with reference to the water stream; 4) manipulating **constrained articulated object** (i.e., turning on and off faucet) requires mechanical compliance provided by soft fingers; 5) the policy also need to be **semantically robust** to the concept of "cleanliness". When additional ketchup is added during washing or even after the washing phase is done, the robot needs to resume washing and wiping.

Performance A single demonstrator collected 258 demonstrations with randomized initial states including, ketchup patterns, position of the plate and sponge, along with water faucet angle. The collected demo also include explicit demonstrations of recovery behavior when additional ketchup is added. For this task, we train diffusion policy by finetuning a CLIP [29] pretrained ViT-B/16 [11] vision encoder. Overall, UMI achieves $14/20 = 70\%$ success rate. In addition, we demonstrate the robustness of our policy against various distractors, and types of sauce (mustard, chocolate syrup, caramel syrup), as well as robustness against perturbations, see Fig. 10 and video on our website for details.

No CLIP-pretrained ViT vision encoder. For this visually complex task, we found training ResNet-34 [17] from

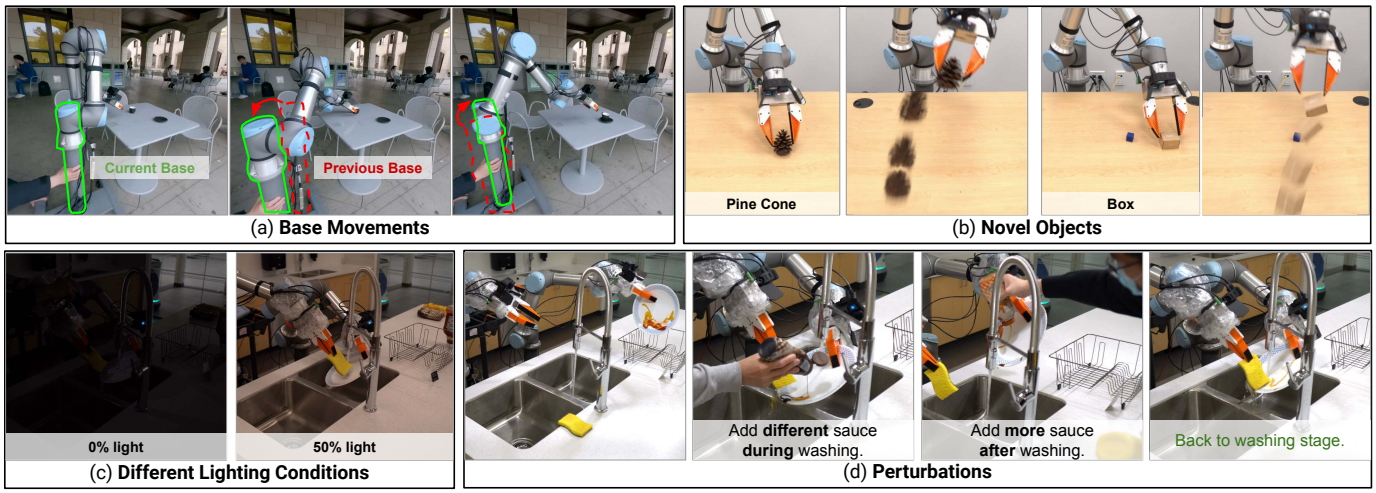


Fig. 10: **Robustness Test.** We test the policy robustness with different inference time perturbations such as moving robot base, novel objects, different lighting conditions, and adding different and more condiments for the dish-washing task. The learned policy is robust against these perturbations and completes the task. Please check the video on our website for more details.

scratch to be insufficient. Specifically, the baseline policy with ResNet-34 learned an non-reactive behavior and ignored any variation in plate or sponge position. As a result, it cannot perform the task, $0/10 = 0\%$.

VI. In-the-Wild Generalization Experiments

Prior works in behavior cloning typically only evaluate in the same environment as data collection, often limited by their inability to collect sufficiently diverse dataset to allow generalization. By not relying on teleoperation with real robots, UMI enables low-cost data collection in any environment, which we refer to as in-the-wild data.

We evaluate UMI's ability to produce generalizable visuomotor policies by scaling up the cup arrangement task Sec. V-A to **novel environments** and **novel objects**. Within 12 person-hours, 3 demonstrators collected 1400 demonstrations for the cup arrangement task across 30 diverse physical locations, including homes, offices, restaurants, and outdoor environments. The demonstrations involved 15 espresso cups of different colors, shapes (cylindrical and tapered), and materials (ceramic, glass, and metal). To ensure model capacity, we increased the vision encoder further to CLIP [30] pretrained ViT-L/14 [11]. We evaluated our UMI policy in two unseen environments (Fig. 9 first column):

- **Cafe table** is a metal table in the outdoor seating area of a busy cafe where a large number of pedestrians serve as natural distractors. We tested 5 cups in the training set and 2 testing (unseen) cups, with 5 initial poses each, 35 experiments in total.
- **Water fountain** is a black cubic water fountain, with a thin film of water constantly flowing from the center, covering the entire top surface. This environment is notably out-of-distribution since all of our demonstrations are collected on non-black tables, not to mention changes in surface dynamics due to the presence of water. We tested 3 training and 2 testing cups, with 5 initial poses each, 25 experiments in total.

We selected the 2 testing cups such that one has an out-of-distribution color (dark blue), while the other has an unseen texture (brown rings). For each test case, we vary the initial pose of both the cup and the saucer.

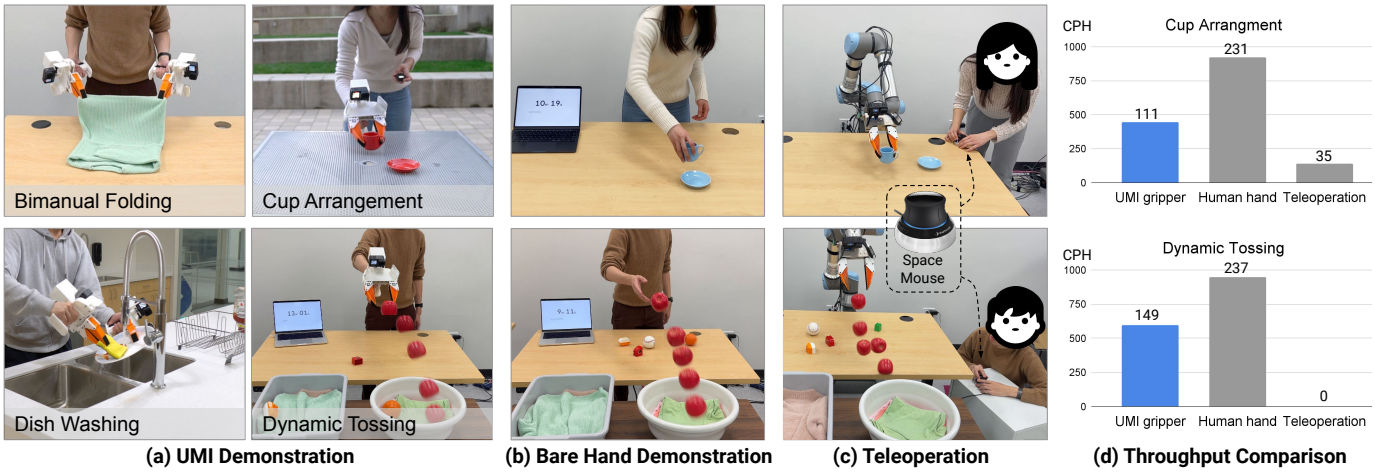
Our UMI policy has $28/40 = 70\%$ success rate on training cups and $15/20 = 75\%$ success rate on testing cups, with a combined success rate of $43/60 = 71.7\%$. More qualitative and quantitative results are shown in Fig. 9.

No in-the-wild data. To validate the generalization ability comes from in-the-wild data, instead of pretrained vision backbone, we trained another model that only uses data from our narrow-domain experiment collected in the same lab environment (described in Sec. V-A) with the same pretrained ViT vision backbone. In the same unseen environments, as shown in Fig. 9(b), the robot with the baseline policy doesn't even move toward the cup. As a result, its success rate is 0%.

Takeaway. This result indicates that finetuning a large pre-trained model with narrow-domain data is insufficient for producing an in-the-wild deployable policy. Therefore collecting diverse, in-the-wild data is still critical for effective generalization to novel environments and objects.

VII. Data Collection Throughput and Accuracy

Throughput. UMI's improved ergonomics and intuitiveness over teleoperations also lead to improved data collection throughput. To demonstrate this effect, we record the number of demonstrations that can be collected within 15 minutes by the same operation using 3 different methods (Fig. 11): 1) Human hand demonstration 2) UMI gripper 3) Spacemouse-based teleoperation, which is a typical teleoperation interface used in many learning from demonstration works [9, 55]. We measure the data throughput on two tasks: 1) cup arrangement 2) dynamic tossing. Note that the time taken to reset the environment, randomize objects, and handle robot faults (such



(a) UMI Demonstration (b) Bare Hand Demonstration (c) Teleoperation (d) Throughput Comparison

Fig. 11: **Data Collection Throughput Comparison** among UMI, bare hand, and teleoperation with a space mouse. UMI is not only faster than traditional teleoperation but also able to perform tasks that were not possible with typical teleoperation interfaces (e.g., dynamic tossing).

SLAM Tracking Mean Error

	Position (mm)	Rotation (degree)
Per-gripper ATE	6.1	3.5
Inter-gripper RPE	10.1	0.8

Fig. 12: **SLAM Accuracy.** We evaluate the SLAM accuracy with a MoCap benchmark including 7 single-gripper tasks and 7 bimanual tasks. Overall, we can achieve less than 1 cm and 4° tracking error.

as self-collisions) are also counted in this experiment to accurately represent the real-world data collection throughput.

On the cup arrangement task, the UMI gripper is more than 3× faster than teleportation, at 48% speed of the human hand, shown in Fig. 11 (d). Note that human is significantly faster on reset and randomization, due to their proximity to the objects. On the dynamic tossing task, the UMI gripper is at 64% speed of the human hand, while the teleportation method failed to produce a single successful demonstration in 15 minutes.

Accuracy. To independently assess the accuracy of our SLAM-based tracking system, we collected a SLAM benchmark dataset with MoCap ground truth. The dataset contains 7 single-gripper tasks and 7 bimanual tasks, all with a variety of movable objects in view as well as natural and rapid human motion. As shown in Fig. 12, our SLAM system has a mean Absolute Trajectory Error (ATE) of 6.1mm for position and 3.5° for rotation. Since both grippers are localized with the same map, we can also obtain the relative pose between two grippers (i.e. inter-gripper pose PD2.3). The mean Relative Pose Error (RPE) between two grippers is 10.1mm for position and 0.8° for rotation.

VIII. Limitations and Future Works

While UMI demonstrates policy efficacy across a wide range of tasks and scenarios, a few limitations remain. First, since the kinematics limits of the downstream deployment robots are unknown at the time of data collection, we rely on data filtering to ensure the kinematic feasibility of the resulting policy. Future works could develop an embodiment-aware policy learning framework that can transfer skills from valid but hardware-infeasible actions.

Second, our SLAM-based action recovery system inherits visual SLAM's requirement for sufficient texture in the environment. Future works could leverage static third-person-view cameras, coupled with additional fiducial markers on UMI grippers to recover action even in texture-deficient environments like rooms with pure white walls.

Third, collecting data with UMI grippers is still less efficient than human hand demonstration, as shown in Sec. VII. This is in part due to the gripper's weight and bulkiness, and in part due to the reduced degrees of freedom compared to human hands. Future works could explore lighter materials and further improve UMI gripper's mechanical design and ergonomics, or alternatively, build sufficiently capable dexterous robotic hands and policies that can directly transfer from human motions.

IX. Conclusion

We present Universal Manipulation Interface (UMI), a framework that enables learning capable and generalizable manipulation policies directly from in-the-wild human demonstrations. The UMI gripper, a hand-held demonstration interface, captures sufficient information to learn some challenging manipulation tasks, including washing a dirty dish, bimanual sweater folding, and dynamic object tossing and sorting. At the same time, UMI remains highly scalable for in-the-wild data collection with its portability, cost-effectiveness, and operational simplicity. By recording all information in a single, standardized MP4 file, UMI's data can be easily shared over the Internet, allowing geographically distributed data collection from a large pool of nonexpert demonstrators. Our goal with UMI is to democratize robotic data collection, fostering a vast, diverse, and decentralized dataset to emerge from the robotics community

Acknowledgments

This work was supported in part by the Toyota Research Institute, NSF Award #2037101, and #2132519. We want to thank Google and TRI for the UR5 robots, and IRIS and IPRL lab for the Franka robot hardware. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

We would like to thank Andy Zeng, Pete Florance, Huy Ha, Yihuai Gao, Samir Gadre, Mandi Zhao, Mengda Xu, Alper Canberk, Kevin Zakka, Zeyi Liu, Dominik Bauer, Tony Zhao, Zipeng Fu and Lucy Shi for their thoughtful discussions. We thank Alex Alspach, Brandan Hathaway, Aimee Goncalves, Phoebe Horgan, and Jarod Wilson for their help on hardware design and prototyping. We thank Naveen Kuppuswamy, Dale McConachie, and Calder Phillips-Graffine for their help on low-level controllers. We thank John Lenard, Frank Michel, Charles Richter, and Xiang Li for their advice on SLAM. We thank Eric Dusel, Nwabisi C., and Letica Priebe Rocha for their help on the MoCap dataset collection. We thank Chen Wang, Zhou Xian, Moo Jin Kim, and Marion Lepert for their assistance with the Franka setup. We especially thank Steffen Urban for his open-source projects on GoPro SLAM and Camera-IMU calibration, and John @ 3D printing world for inspiration of the gripper mechanism.

References

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[4] Anais Brygo, Ioannis Sarakoglou, Nadia Garcia-Hernandez, and Nikolaos Tsagarakis. Humanoid robot teleoperation with vibrotactile based balancing feedback. In *Haptics: Neuroscience, Devices, Modeling, and Applications: 9th International Conference, EuroHaptics 2014, Versailles, France, June 24-26, 2014, Proceedings, Part II 9*, pages 266–275. Springer, 2014.

[5] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015. doi: 10.1109/ICAR.2015.7251504.

[6] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[7] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. doi: 10.1109/TRO.2021.3075644.

[8] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from "in-the-wild" human videos. In *Proceedings of Robotics: Science and Systems (RSS)*, 2021.

[9] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[10] Kiran Doshi, Yijiang Huang, and Stelian Coros. On hand-held grippers and the morphological gap in human manipulation demonstration. *arXiv preprint arXiv:2311.01832*, 2023.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] Jiafei Duan, Yi Ru Wang, Mohit Shridhar, Dieter Fox, and Ranjay Krishna. Ar2-d2: Training a robot without a robot. *arXiv preprint arXiv:2306.13818*, 2023.

[13] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[14] Hongjie Fang, Hao-Shu Fang, Yiming Wang, Jieji Ren, Jingjing Chen, Ruo Zhang, Weiming Wang, and Cewu Lu. Low-cost exoskeletons for learning whole-arm manipulation in the wild. *arXiv preprint arXiv:2309.14975*, 2023.

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

[16] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2014.01.005. URL https://www.sciencedirect.com/science/article/pii/S0031320314000235.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr

abs/1512.03385 (2015), 2015.

[18] GoPro Inc. Gpmf introuction: Parser for gpmf™ formatted telemetry data used within gopro® cameras. https://gopro.github.io/gpmf-parser/. Accesssed: 2023-01-31.

[19] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, volume 164, pages 991–1002. PMLR, 2022.

[20] Moo Jin Kim, Jiajun Wu, and Chelsea Finn. Giving robots a hand: Broadening generalization via hand-centric human video demonstrations. In *Deep Reinforcement Learning Workshop NeurIPS*, 2022.

[21] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

[22] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning (CoRL)*, volume 87, pages 879–893. PMLR, 2018.

[23] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 892–909. PMLR, 2022.

[24] Chuer Pan, Brian Okorn, Harry Zhang, Ben Eisner, and David Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 1783–1792. PMLR, 2023.

[25] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[26] Luka Peternel and Jan Babič. Learning of compliant human–robot interaction using full-body haptic interface. *Advanced Robotics*, 27(13):1003–1012, 2013.

[27] Pragathi Praveena, Guru Subramani, Bilge Mutlu, and Michael Gleicher. Characterizing input methods for human-to-robot demonstrations. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 344–353. IEEE, 2019.

[28] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3:297–330, 2020.

[31] Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task-agnostic offline reinforcement learning. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 1838–1849. PMLR, 2023.

[32] Felipe Sanches, Geng Gao, Nathan Elangovan, Ricardo V Godoy, Jayden Chapman, Ke Wang, Patrick Jarvis, and Minas Liarokapis. Scalable. intuitive human to robot skill transfer with wearable human machine interfaces: On complex, dexterous tasks. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6318–6325. IEEE, 2023.

[33] Karl Schmeckpeper, Annie Xie, Oleh Rybkin, Stephen Tian, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*, pages 708–725. Springer, 2020.

[34] Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. In *Proceedings of the 2020 Conference on Robot Learning (CoRL)*, volume 155, pages 339–354. PMLR, 2021.

[35] Mingyo Seo, Steve Han, Kyutae Sim, Seung Hyeon Bang, Carlos Gonzalez, Luis Sentis, and Yuke Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2023.

[36] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[37] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[38] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 654–665. PMLR, 2023.

[39] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *Proceedings of The 7th Conference on Robot Learning (CoRL)*, volume 229, pages 405–424. PMLR, 2023.

[40] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields:

Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022.

[41] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *Robotics and Automation Letters*, 2020.

[42] H.J. Terry Suh, Naveen Kuppuswamy, Tao Pang, Paul Mitiguy, Alex Alspach, and Russ Tedrake. SEED: Series elastic end effectors in 6d for visuotactile tool use. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4684–4691, 2022. doi: 10.1109/IROS47612.2022.9982092.

[43] Alexander Toedtheide, Xiao Chen, Hamid Sadeghian, Abdeldjallil Naceri, and Sami Haddadin. A force-sensitive exoskeleton for teleoperation: An application in elderly care robotics. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12624–12630. IEEE, 2023.

[44] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Proceedings of The 7th Conference on Robot Learning (CoRL)*, volume 229, pages 201–221. PMLR, 2023.

[45] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[46] Josiah Wong, Albert Tung, Andrey Kurenkov, Ajay Mandlekar, Li Fei-Fei, Silvio Savarese, and Roberto Martín-Martín. Error-aware imitation learning from teleoperation data for mobile manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, volume 164, pages 1367–1378. PMLR, 2022.

[47] Philipp Wu, Fred Shentu, Xingyu Lin, and Pieter Abbeel. GELLO: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023*, 2023.

[48] Keenan A Wyrobek, Eric H Berger, HF Machiel Van der Loos, and J Kenneth Salisbury. Towards a personal robotics development platform: Rationale and design of an intrinsically safe personal robot. In *2008 IEEE International Conference on Robotics and Automation*, pages 2165–2170. IEEE, 2008.

[49] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv:2203.06173*, 2022.

[50] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.

[51] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. In *Conference on Robot Learning (CoRL)*, volume 155, pages 1992–2005. PMLR, 2021.

[52] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635. IEEE, 2018.

[53] Zichao Zhang, Henri Rebecq, Christian Forster, and Davide Scaramuzza. Benefit of large field-of-view cameras for visual odometry. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 801–808, 2016. doi: 10.1109/ICRA.2016.7487210.

[54] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[55] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205, pages 1199–1210. PMLR, 2023.