Enhancing Robot Learning from Demonstration through Mixed Reality Capture and Interventions

Research Proposal

Jiahao Chen

Student ID: 1118749

Supervisors:

Wafa Johal

Qiushi Zhou

Master of Computer Science

School of Computing and Information Systems

The University of Melbourne

September, 2023

Enhancing Robot Learning from Demonstration through Mixed Reality Capture and Interventions

Abstract

Learning from Demonstration (LfD) plays a crucial role in human-robot collaboration, allowing humans to directly impart desired behaviours to robots. A significant challenge is bridging the information gap between humans and robots due to the difficulty in understanding the robot's learning status and the absence of methods for evaluating a human's perception of this status. With the recent rise of mixed reality (XR) headsets, innovative human-robot interaction (HRI) techniques using XR have emerged. XR offers advantages over traditional interaction modalities, such as superimposing holographic entities onto the real world, bypassing the constraints of the physical environment, and delivering a more immersive user experience. Meanwhile, integrating XR with LfD remains a novel field worthy of exploration. Our study introduces a novel approach that utilises the Mixed Reality Capture (MRC) of the Microsoft Hololens 2 in tandem with object detection techniques. Unlike traditional kinesthetic teaching or teleoperation, which requires expertise in robot manipulation, MRC enables even novice users to directly demonstrate desired behaviors with their hands, without the need to physically touch the robot or be in close proximity. Moreover, we aim to design and implement a range of innovative user interface (UI) interventions, aiming to enhance the teaching process. These tools offer users multiple abilities, including revising prior demonstrations, viewing the robot's intended trajectories, checking various demonstration levels, etc. A sorting task will serve as a case study to determine the influence of these UI interventions on the teaching efficiency.

Contents

1 Introduction

Autonomous robots have been developed and applied to the industrial field to enhance productivity, including sectors such as manufacturing, supply chain management, construction, etc. However, proficiency in programming skills for robots is imperative when it comes to enabling robots to learn new tasks, a process that is also significantly time-consuming [1]. In this scenario, several existing works are focused on human-robot interaction (HRI), aiming to facilitate interactions between robots and users with limited knowledge of programming and robotics [2, 3]. Learning from Demonstration (LfD) is one of the main elements in the domain of human-robot collaboration [2], representing a mechanism through which robots can directly acquire skills by observing or emulating humans' executions [1, 2]. Experts can effectively design and customise robot behaviours without the need for learning in-depth robotics knowledge, highlighting the huge potential of LfD in industries such as manufacturing and health care [1].

With the widespread of virtual, augmented, and mixed reality (VAM) headsets (e.g. Microsoft Hololens, Meta 2, Magic Leap, HTC Vive, Oculus Quest, Apple Vision Pro, etc) at the consumer level, novel methods of HRI utilising such technologies are recently being explored [4, 5]. Mixed reality (XR), particularly, can be regarded as the fusion of virtual reality (VR) and augmented reality (AR) [6], allowing users to interact with virtual objects placed in the real world. In contrast to conventional interactions that mainly depend on the robot's internal physical or visual feedback mechanisms (e.g. robot's movements, gestures or gaze outputs, physical transformation, signal lights, small displays, etc), XR interfaces can be designed without the constraints of the physical world or robot's physical design [7], additionally providing users with 3D visualisations [5]. VAM technologies can be integrated with robotics across a variety of application domains, and one of the most common areas in prior research involves the utilisation of VAM headsets to perform remote robot teleoperation [5]. For instance, Zollmann et al. [8] introduce an AR interface for controlling unmanned aerial vehicles , displaying virtual spheres and shadows to indicate robot's information (e.g. way points, altitude, trajectory, etc) for human supervision and intervention. However, the application of XR for robot LfD remains relatively novel, making it a promising area worthy of in-depth exploration.

A significant challenge in LfD arises from the barriers present in human-robot communication, given the difficulty of understanding and predicting robot intentions for humans [5]. Walker et al. [5] suggest that the problem is due to the inadequate exchange of information between robots and humans, akin to the model proposed by Pea,

which characterises this as the *Gulf of Execution* and *Gulf of Evaluation* [9]. The Gulf of Execution denotes the challenge of converting high-level commands from humans into understandable instructions for robots [5]. On the other hand, the Gulf of Evaluation represents the difficulty that robots encounter in generating useful feedback for humans to effectively assess their state [5]. Phaijit et al. [10] also emphasise the importance of enabling the human teacher to have a comprehensive understanding of the robot's learning status and mental model in the process of LfD.

Another challenge is the correspondence problem stemmed from visual demonstrations, a technique that is user-friendly especially for non-roboticists as it does not require direct physical interaction with robots [1]. Differences in the demonstrator's actuation space compared to that of the robot can lead to variances in dimensionality and motion constraints between the two systems [1]. XR headsets potentially offers a solution to this problem. For example, leveraging the Mixed Reality Capture (MRC) [11] capabilities of devices such as Hololens enables robots to directly learn from the user's perspective without mapping procedures. Additionally, including virtual objects in the MRC can possibly eliminate physical limitations.

Previous studies involve integration of VAM to improve kinestheic LfD, a type of LfD method that relies solely on the development of the robot's hardware without the requirement of additional sensors or interfaces [1]. For example, Luebbers et al. [4] present an AR interface designed for the visualisation and manipulation of robot LfD based on the Concept-Constrained LfD proposed by Mueller et al [12]. The interface visualises trajectories demonstrated by humans or learned by robots, together with the overlaid constraints within the physical environment. In addition, Phaijit et al. [10] propose a couple of potential designs for user interface (UI) interventions aimed at improving generic robot LfD. They investigate two of the interventions (visualizing decision trees learned by robots and visualising demonstation levels), proving their ability to reduce the number of demonstrations and increase the interpretability of robot's mental model.

In this research, we plan to conduct two studies in order to address two research questions outlined below, stemming from the challenges mentioned earlier. Figure 1 illustrates the flow of all phases and their respective outcomes.

- **RQ1**: *How can we leverage MRC to assist in the creation of visual demonstrations?*
- **RQ2**: *Can XR interventions enhance the teaching efficiency for users lacking expertise in robotics?*

In the initial stage (Phase 0), we will establish the foundation for our studies. Given that our primary focus does not lie in the control theory, we have chosen to conduct our

studies using a sorting task. In this task, users are required to teach the robotic arm to classify objects with different colors or shapes into two distinct categories, with the sorting rules defined by the user themselves. The *robomimic* framework, as introduced by Mandlekar et al. [13], will be implemented. This framework provides various datasets and learning algorithms for different tasks in robot LfD. We will first collect new data for our sorting task through teleoperation and train the robotic arm to be capable of picking and placing objects.

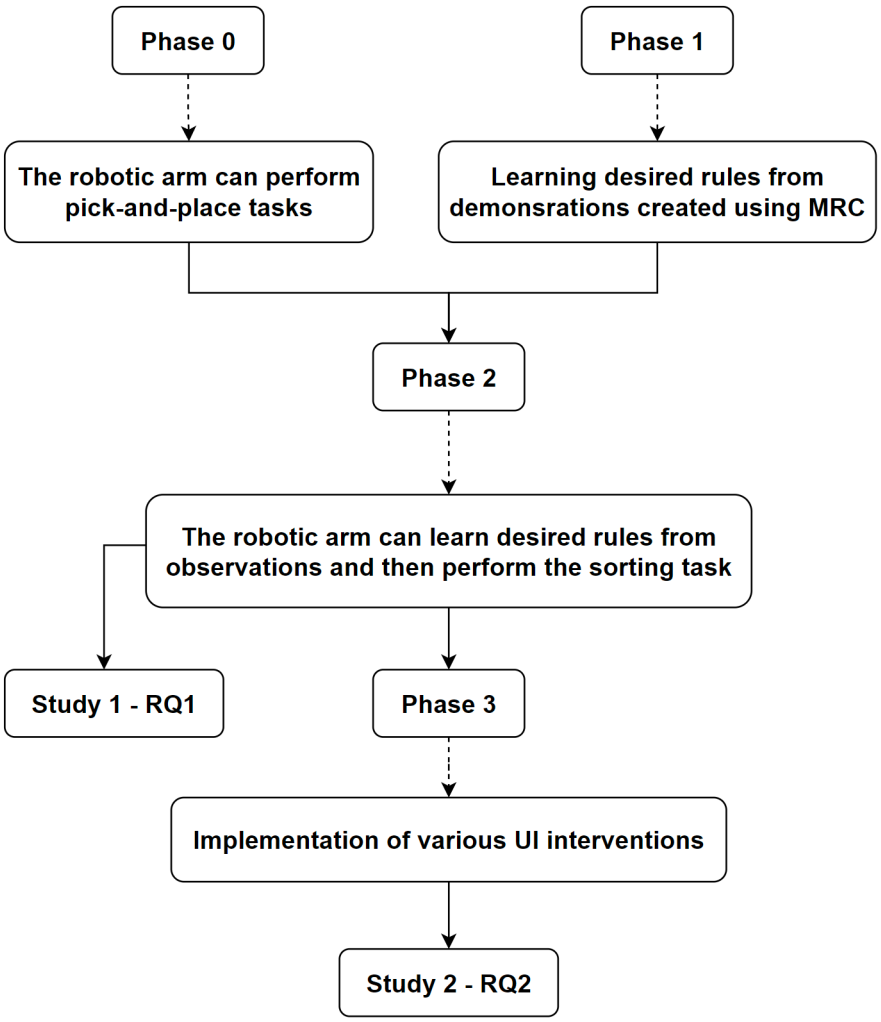

Figure 1: Flow chart of phases and outcomes

In Study 1, our aim is to investigate whether the integration of MRC can assist users in creating visual demonstrations for the robotic arm. The users will place two virtual regions in Hololens 2 and sort objects into different regions using their own

hands. Study 1 comprises two phases. In Phase 1, we will initially apply various object detection algorithms to identify locations of objects within respective regions. Next, we will translate the detection outcomes into training data, which will serve as input for learning the user-defined rules. Moving on to Phase 2, we will merge the outcomes from both Phase 0 and Phase 1 by developing an application that connects XR with the robotic arm. We will evaluate its performance in the sorting task with the objective of answering RQ1.

In Study 2, we will implement a couple of UI interventions inspired by the concepts introduced by Phaijit et al. [10], such as editing demonstration, visualisation of robot's range of motion, presenting demonstration levels for each object, etc. Moreover, we will explore methods of enhancing users' understanding of the robot's learning status. Finally, a user study will be conducted to assess the effect of these interventions on teaching efficiency, thus addressing RQ2. Figure 2 presents an envisioned picture of the HRI scenario within this task.

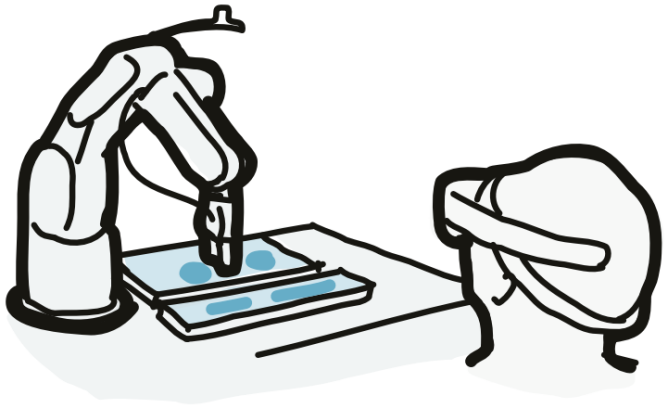

Figure 2: Envisioned HRI scenario [7]

In summary, our contributions are listed as follows:

- We will leverage XR capabilities, especially MRC, together with computer vision to assist novice users in creating visual demonstrations for robots.
- The implementation of novel designs for XR-based UI interventions, with the aim of enhancing the efficiency of users without expertise in robotics in teaching robots.
- A user study that investigates the impact of various UI interventions designed to improve the efficiency of robot LfD.

2 Literature Review

This section will be divided into two parts, (1) XR in Robotics and (2) Robot LfD. The objective of the first part is to establish an overview of the current applications of XR in robotics. The second part aims to summarise various approaches in robot LfD.

2.1 XR in Robotics

2.1.1 Defining AR and XR

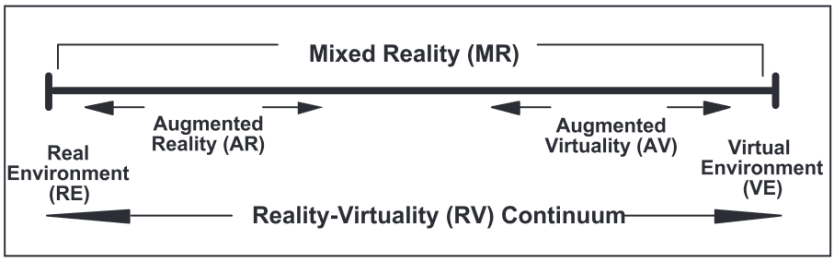

Figure 3: Reality-Virtuality Continuum defined by Milgram et al. [14]

The integration of VAM technologies into robotics has emerged as a novel research field, driven by advancements in VAM headsets. Unlike VR, for which there is a general consensus defining it as creating a completely virtual world [6], AR and XR have various definitions, giving rise to certain ambiguities. Azuma [15] provides definitions of AR that includes three characteristics: "*(1) combines real and virtual, (2) interactive in real time, (3) registered in 3-D*". Milgram et al. [14] introduce the concept of "*reality-virtual continuum*" spectrum as figure 3 shows, within which XR includes the entire range, spanning from the real environment to augmented reality, augmented virtuality, and virtual environment. This concept remains one of the most prevailing definitions for XR nowadays. In addition, Speicher et al. [6] summarise definitions of XR into four categories: "*(1) XR according to the Reality–Virtuality Continuum, (2) XR as a Combination of AR and VR, (3) XR as 'strong' AR, (4) XR as a synonym for AR*", and suggest the definition of XR can vary based on different contexts. Recently, XR has achieved more attention than AR, possibly because of the release of Apple Vision Pro, indicating its potential for increased adoption in the future. In this study, we define XR as the fusion of AR and VR. Our primary focus will be on XR headsets, rather than on other devices such as tablets or projectors. Note that certain literature may use AR in a broader context that includes XR features, and it is important to take these viewpoints into consideration as well.

2.1.2 Forms of XR integration in Robotics

Previous studies involve the establishment of taxonomies for this integration, consisting of aspects such as application domains, design strategies, interaction modalities, etc. In this section, we will first introduce diverse forms of XR integration in robotics, which offer possibilities for novel forms of HRI.

Phaijit et al. [16] propose a categorisation of different types of HRI nowadays, comprising of "*(1) kinesthetic interaction, (2) graphical user interface, (3) teleoperation, (4) Internet of Things (IoT) mediated, (5) simulation, (6) VR, (7) XR*". They point out the importance of shaping the interaction and relationships among humans, robots, virtual content and the physical world. Additionally, they classify AR in robotics into augmented human perception and augmented robot perception, suggesting that these also can be combined to enhance the shared perception of users and robots. Suzuki et al. [7] suggest that methods of AR in robotics can be classified according to their location, including on-body, on-robot, and on-environment. Here, we present a summary of the categorisation based on previous studies, as shown below:

- **Augmenting Robots** [7, 16]: Robots are overlaid with additional information in XR devices, such as adding cartoon faces onto a robot to convey expressive emotions [17], or displaying the motion intent of a robotic arm [18].
- **Augmenting Environments** [7, 16]: Augment the surrounding scene, including mid-air space, real objects, physical environments, etc. An example of this is presenting information in the surroundings of robots by using a large surface [19].
- **Augmenting Interactive Objects** [16]: Both the user and the robot can interact with virtual objects positioned in the physical world. For instance, instead of using real objects, virtual replicas can be rendered in manipulation tasks [20].
- **Augmenting User Interfaces** [16]: Virtual entities are anchored to the users, ignoring their orientation or position. This can be applied as a virtual monitor, which can display the robot's camera perspective during teleoperation [21].

Within each category, there exist numerous design variations, and these categories can also be combined with other technologies. For example, Yamamoto et al. [22] employ AR in robot-assisted surgery, integrating technologies such as image processing and 3D surface reconstructions. In this study, we will investigate combination of techniques like computer vision with XR capabilities to improve robot LfD.

2.1.3 Multimodal Interaction

While XR merges the virtual and real worlds visually, it also provides multimodal interfaces that open up new possibilities in HRI. Following is a summary of the current interaction modalities in XR with robotics, provided by Suzuki et al. [7]:

- **Tangible**: Users can directly interact with physical objects. For instance, kinesthetic LfD requires users to manipulating the robot through the their desired motions [1]. Note that the use of XR headsets is not compulsory in this approach.
- **Touch**: Users can manipulate touch screens to interact with robots, with touch-based interaction providing more precise control (e.g. robot's motion) [7]. Virtual screens can be created in XR to facilitate HRI.
- **Pointer**: Pointers serve as a form of spatial interaction method, which allow users to implicitly communicate with robots.
- **Controller**: Controllers offer users a means of explicit communication with robots, including haptic or vibration feedback, which can increase manipulation efficiency.
- **Gesture**: In XR headsets, gestures are frequently used as a spatial interaction method. For example, in Hololens 2, users can use various gestures to perform operations such as selection, dragging, zooming, etc.
- **Gaze**: Gaze is usually combined with gestures to perform actions. It can be divided into head-tracking and eye-tracking, with head-tracking generally being more reliable in the past [23]. However, in Hololens 2, the accuracy of eye-tracking has been significantly increased by the eye-tracking camera, paving the way for new levels of gaze interaction [24].
- **Voice**: Voice input can be used to issue commands, particularly when users and robots are located in the same environment.
- **Proximity**: Robots can respond to the location and behavior of humans. For example, when a passerby is detected, robots can display their intended motion to ensure safety [25].

The fusion of multiple interaction modalities is a promising direction for future studies [7]. Hololens 2 supports several types of multimodal interactions with virtual entities, mainly through a combination of gestures, gaze and voice commands. For example, the

user can gaze at a hologram and say "put this", and then look at where s/he want to place it and say "over here" [26]. Such kind of interaction can also be extended to HRI. As an example, Krupke et al. [23] investigate the performance of two approaches for co-located HRI using XR headset (Hololens 1), including (1) head orientation and (2) pointing, both of which are combined with voice commands. Their experiments indicate that heading-based approaches are more accurate and efficient in pick-and-place tasks. However, their findings are limited to Hololens 1, and further investigation is necessary to assess performance on later devices, such as Hololens 2.

In addition, with the progress in Large Language Models (LLM) such as ChatGPT, recent research has been dedicated to the integration of LLMs with robotics, enabling robots to understand natural language commands. For example Huang et al. [27] have proposed VLMaps, a method that maps the physical environment using visual-language features so that mobile robots can be navigated using natural languages. Similarly, Brohan et al. [28] have introduced RT-2, a model that has great generalisation capabilities and can utilise multi-stage semantic reasoning to solve complex tasks. Hypothetically, these advancements can be combined with voice input to create innovative methods of HRI. Moreover, LLMs have the potential to bridge the information gap between humans and robots as it uses natural languages, which especially benefits novice users.

2.2 Robot LfD

LfD, also known as imitation learning or behavioural cloning, has gained growing research interest over the past decade [1]. As previously mentioned, LfD offers a way for users lacking expertise in robotics to program robots. Ravichandar et al. [1] suggest that LfD can be regarded as a supervised-learning problem because it relies on demonstrations provided by human teachers, and outline several challenges in LfD. It faces challenges stemmed from both machine-learning (e.g. the curse of dimensionality, noisy data, sparse datasets, etc) and control theory (e.g. external disturbances, convergence, stability, etc). Moreover, existing challenges from HRI, for example, variability in human performance and knowledge among various human subjects, need to be addressed. In this research, our main focus will be on leveraging XR capabilities to deal with the challenges that arise in HRI.

2.2.1 Categorisation

LfD can be divided into three categories based on types of demonstrations, as the following shows [1]. Figure 4 provides examples of each category.

Figure 4: Examples of three types of robot LdD provided by Ravichandar et al. [1]

- **Kinesthetic Teaching**: Users create demonstrations by directly manipulating the robot to execute the intended motions for a variety of tasks [29]. The robot's sensors capture several states, such as joint angles and torques. Therefore, it does not face correspondence problems because the recorded states can be directly used as training data. The quality of demonstrations can affect the learning outcomes, which requires expertise of manipulators and make it less user-friendly for novice users. Furthermore, implementing this approach on platforms such as robotic hands or legged robots is more difficult than on robotic arms.

- **Teleoperation**: Robots are manipulated using external devices, such as controllers and VAM interfaces [30]. In comparison to kinesthetic teaching, teleoperation does not require users to be physically co-located with the robots, thereby allowing remote demonstrations. Additionally, it can be applied to more complex platforms like humanoids [30] and robotic hands [31]. However, teleoperation requires extra effort to develop a control interface for the robot. Zhu et al. [32] have introduced *robosuite*, a framework that offers a variety of controller implementations for multiple robots.

- **Observation**: In contrast to participating in the process of demonstration, robots passively observe the user's actions in this approach [33]. The user presents the desired task through their body movements or by wearing additional sensors for tracking, eliminating the need for users to have expertise in robot manipulation. However, this method faces several challenges, for instance, the correspondence problem due to differences in the demonstrator's and the robot's locations. Moreover, it needs to encode and map human actions into a format understandable by robots, while there may exist more noise in data generated by human.

2.2.2 Formalisation of LfD

Billing et al. [34] have provided a unified formalisation of the concepts in robot LfD as presented in figure 5:

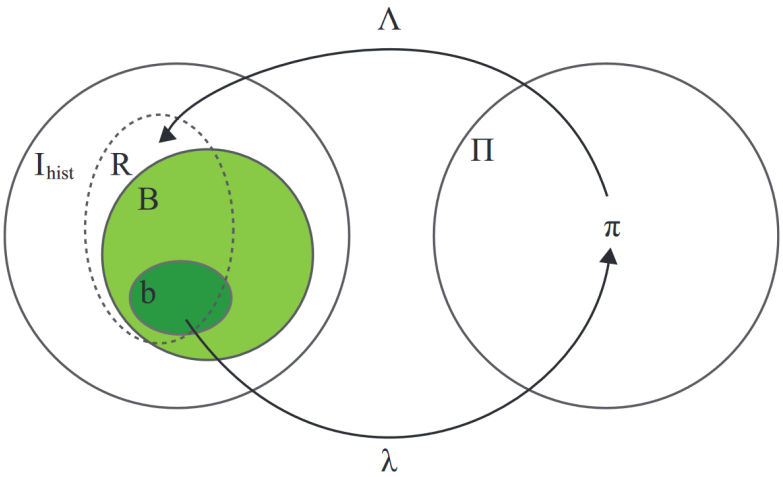

Figure 5: Formalisation of the LfD defined by Billing et al. [34]

- **The wanted behaviour (B)**: B represents the behaviours that the human teacher expects the robot to acquire from demonstrations.
- **Controller space (Π)**: The controller space includes all potential controllers designed for the observation and action spaces relevant to the desired task.
- **Information space (I)**: The observation and action spaces, also referred to as the sensory-motor space [35], can be denoted as I. Specifically, I_{hist} involves all combinations of the robot's historical observations and actions.
- **Demonstration set (b)**: The human teacher provides the robot with N demonstrations to illustrate the desired task, denoted as follows. Each single demonstration is represented by $\beta^{(n)}$.

$$b = \{\beta^{(1)}, ..., \beta^{(N)}\} \subset B \tag{1}$$

- **Learning function (λ)**: The process of a robot learning desired behaviour (B) from a collection of demonstrations (b) is interpreted as the selecting π from the controller space Π, utilising the learning function λ.

$$\pi = \lambda(b) \in \Pi \tag{2}$$

- **Realisation function (Λ)**: The realisation function is employed to generate a realisation space (R) that maps the chosen controller to a task previously executed by the robot from its historical observations and actions (I_{hist}).

$$R = \Lambda(\pi) \in I_{hist} \tag{3}$$

Based on this formalisation, Sena et al. [36] have proposed a concept to quantify human teaching behaviour as figure 6 depicts.

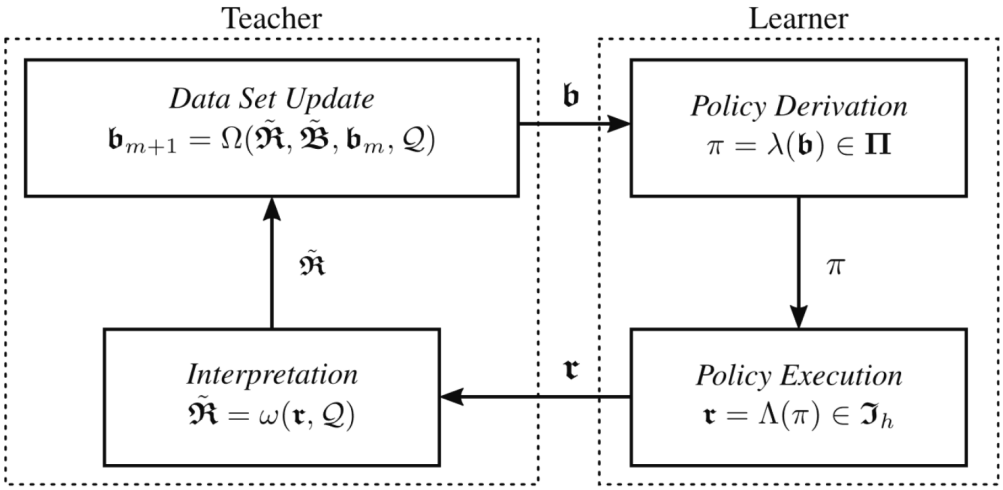

Figure 6: The LfD pipeline proposed by Sena et al. [36]

The belief space (M) is introduced to model the human teacher's understanding of the robot's learning progress. The teacher estimates the robot's learning status based on observed realizations (R) as well as latent factors (Q) such as teacher's proficiency in the task or mental state (teacher's bias). The interpretation function (ω), as shown below, maps R and Q to the teacher's estimation ($\tilde{R}$).

$$\tilde{R} = \omega(R, Q) \subset M \tag{4}$$

Subsequently, the teacher intends to create new set of demonstrations based on the estimated learning status ($\tilde{R}$) and his/her understanding of the desired behaviour ($\tilde{B}$). The demonstration function (Ω) is defined as follows to model the new demonstration set (b_{m+1}) according to the current status.

$$b_{m+1} = \Omega(\tilde{R}, \tilde{B}, b_m, Q) \tag{5}$$

They have also proposed methods for assessing teaching performance. The teaching efficacy (ε) is used to evaluate the accuracy of the robot in executing the required task. As the intersection of the realisation space (R) and the target behaviour (B) increases, the robot performs more accurately. Since the robot's interpretation of B may differ from that of the teacher's, there could be policies that have been learned outside of B ($R \setminus B$). These policies are disregarded when calculating ε.

$$\varepsilon = \frac{|R \cap B|}{B}, \varepsilon \in [0, 1] \tag{6}$$

Based on the teaching efficacy, teaching efficiency (η) assesses the overall performance of the LfD process by normalising ε with the number of demonstrations created (—b—). To achieve high teaching efficiency, it is essential to maintain a high level of efficacy while minimizing the number of required demonstrations.

$$\eta = \frac{\varepsilon}{|b|}, \eta \in [0, 1] \tag{7}$$

Phaijit et al. [10] have further extended this formalisation to quantitatively evaluate the accuracy of teacher's interpretation of the robot's learning status, denoted as i_a.

$$i_a = \frac{|R \cap \tilde{R} \cap B| + |B \setminus (R \cup \tilde{R})|}{|B|}, i_a \in [0, 1] \tag{8}$$

- $R \cap \tilde{R} \cap B$ **(true positives)** denotes tasks that the robot executes correctly in both the realisation space (R) and the teacher estimation ($\tilde{R}$).
- $B \setminus (R \cup \tilde{R})$ **(true negatives)** denotes tasks that cannot execute correctly by the robot in both the realisation space (R) and the teacher estimation ($\tilde{R}$).
- $\tilde{R} \cap B \setminus R$ **(false positives)** denotes tasks that cannot execute correctly by the robot, but the user expects them to be learned.
- $R \cap B \setminus \tilde{R}$ **(false negatives)** denotes tasks that the robot executes correctly, but are beyond the teacher's expectations.

2.2.3 XR in Robot LfD

The integration of XR techniques into robot LfD still represents a relatively novel and emerging field, possibly due to the poor performance of XR devices in the past, compared to using traditional devices like keyboard and monitor. Previous research involves

the exploration of combining XR with kinesthetic teaching. In 2019, Luebbers et al. [4] proposed a system employing Hololens 1 for visualizing constraints, such as height and rotation, during creating trajectory demonstrations. In 2021, they further extended the system and introduced ARC-LfD [3], an AR interface designed to assist users in kinesthetic teaching. The interface visualizes the robot's acquired skills, in-situ behaviors, and constraints, while enabling users to verify or modify the learned skills. They also carried out 3 case studies, which presented that ARC-LdF is robust to changes in the task or the environment.

An effective approach for enhancing robot LfD involves implementing UI interventions that enable teachers to receive real-time feedback from the robot regarding its learning status and make adjustments. In 2020, Diehl et al. [37] conducted a study on investigating the effect of three different visualisation techniques developed for a teaching system, including Hololens 1, a tablet with AR simulation and a tablet with RViz (a robot simulation tool). The system includes an AR interface that enables the teacher to inspect the taught behaviour before execution. The interface overlays information on the physical world and provides a semantic explanation for each action. Their study revealed that users preferred AR simulation over RViz due to its ability to provide a more immersive real-world experience. However, they also noted that issues with Hololens 1, such as wearability and a limited field of view (FOV), negatively impacted the user experience. They speculated that Hololens 2 could potentially alleviate these problems.

Recently, Phaijit et al. [10] have introduced a variety of UI interventions based on the formalisation presented above (Section 2.2.2). They suggest that interventions can be applied to four aspects: (1) b (demonstration set), (2) R (observed task realisation), (3) Q (teacher's bias) and (4) B (teacher's interpretation of the desired behaviour). They have implemented two types of interventions in Hololens 2: one displays the decision tree and the other shows the demonstration level of each object. Furthermore, they conducted a user study on a sorting task to evaluate whether these interventions improved teaching efficiency in comparison to giving no feedback. Their findings lead to the conclusion that these interventions can enhance overall efficiency. However, they also noticed that users may struggle to interpret the decision tree, highlighting the importance of exploring more explainable display options.

3 Methodology

The primary focus of this research is to delve into the following research questions:

- **RQ1**: *How can we leverage MRC to assist in the creation of visual demonstrations?*
  Hypothesis 1: *MRC can help novice users to demonstrate with their own bodies.*
- **RQ2**: *Can XR interventions enhance the teaching efficiency for users lacking expertise in robotics?*
  Hypothesis 2: *XR interventions can improve the teaching efficiency.*

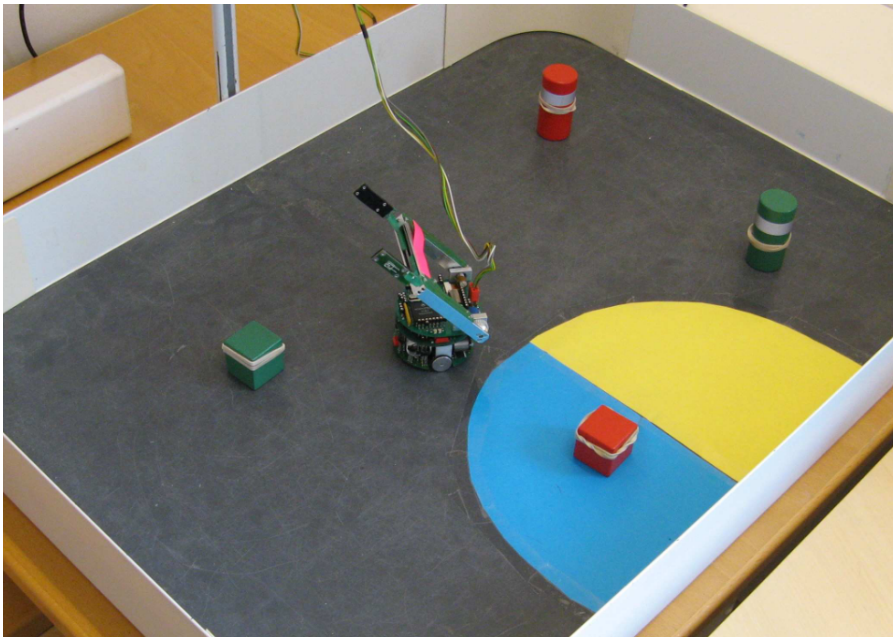

Figure 7: A prototype of sorting task [34]

To investigate the research questions, we intend to establish a case study by extending the sorting task, which is inspired from previous studies [10, 34] as figure 7 presents. Initially, users have the option to establish their own sorting criteria for the objects in front of them and are instructed to wear Hololens 2. Afterward, in contrast to the traditional approach of having colour paints on the robot's platform, users can create two virtual coloured regions overlaid on their own desk, representing two distinct categories. It is worth noting that users are not required to be physically present at the same desk as the robotic arm. Subsequently, they can demonstrate to the robotic arm by placing objects into these different virtual regions. The number of demonstrations is decided by users themselves. The robotic arm then observes these demonstrations and learns the

user-defined sorting rules. Once the user believes they have provided a sufficient number of demonstrations, they can instruct the robot to either sort all the objects or specify which ones to sort.

The research is structured into the following four phases:

- **Phase 0**: In this initial stage, our primary focus is on training the robot to perform pick-and-place tasks, which will serve as a foundational step supporting the subsequent phases.
- **Phase 1**: We aim to integrate MRC and computer vision techniques to enable the robot to learn sorting rules through user demonstrations.
- **Phase 2**: Bridging the robotic arm and Hololens 2, this stage is dedicated to evaluating the system's performance, which addresses RQ1.
- **Phase 3**: The final phase involves the implementation of proposed UI interventions on Hololens 2. We will conduct a user study to assess the resulting improvement in teaching efficiency, providing answers to RQ2.

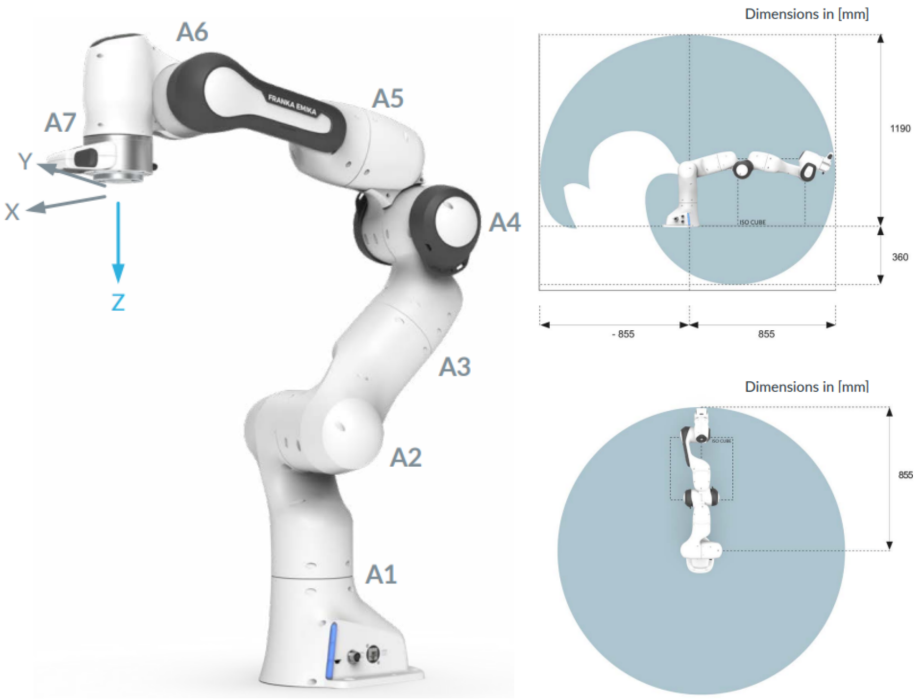

Figure 8: Sensors and reachable space of Franka Emika Panda [38]

3.1 Equipment

This section will first provide an introduction to the equipment utilized in this study, involving both hardware and software components.

3.1.1 Hardware

We will employ the Franka Emika Panda robotic arm for executing pick-and-place tasks. This robotic arm features seven degrees of freedom and is equipped with torque sensors in all seven axes to capture motion status. Moreover, it is adaptable to various accessories, such as grippers for picking up objects and cameras for monitoring the gripper's status or the surrounding environment. Figure 8 shows the distribution of sensors and the reachable space of Franka Emika Panda. In the pick-and-place tasks, we will use either 3D printed cubes or LEGO blocks, and two bins positioned on the robot's platform.

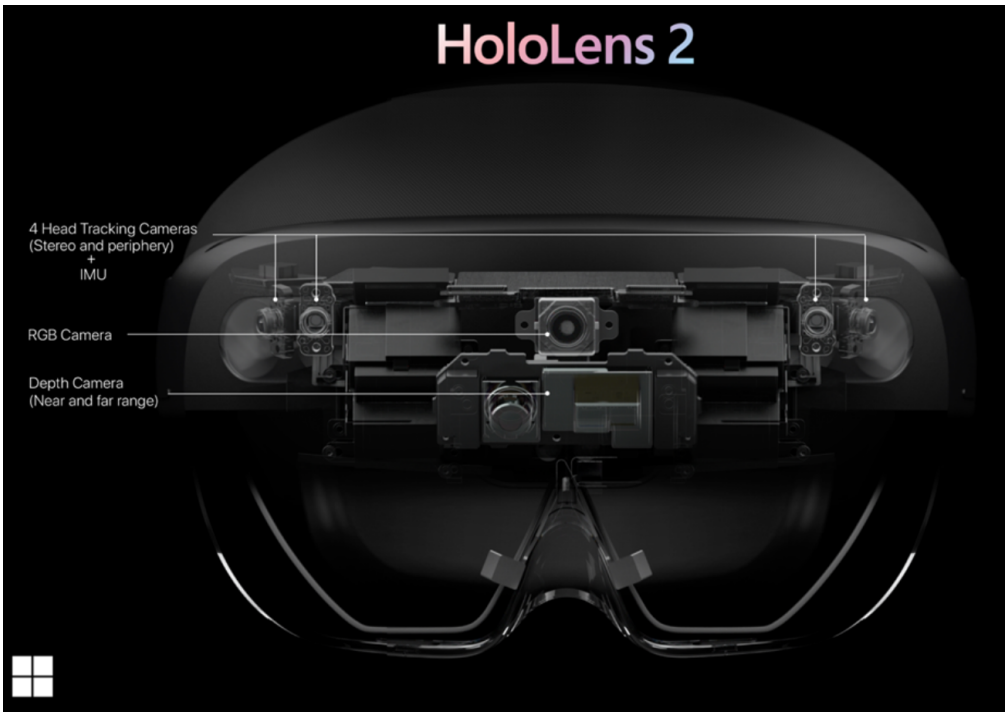

Figure 9: Key components of Hololens 2 [39]

In our XR setup, we will utilise the Microsoft Hololens 2, an untethered XR headset powered by the Windows Holographic OS [39]. This device is equipped with see-through holographic lenses to display virtual entities and offers a range of features, including hand-tracking, eye-tracking, voice commands, MRC, etc. Notably, compared to its pre-

decessor (Hololens 1), Hololens 2 has a larger FOV and a more ergonomic design, which significantly improves the overall user experience. Figure 9 presents the key components of Hololens 2, supporting the mentioned features.

3.1.2 Software

The Robot Operating System (ROS) is an open-source middleware that facilitates the development of robot applications, which offers a variety of software libraries and tools. Franka Emika have provided a Franka Control Interface (FCI) named libfranka [40], which encapsulates the low-level control of the Franka Emika Panda. In this research, the development of the robotic arm will be mainly based on ROS 2 Humble [41] and libfranka. Furthermore, it is typically necessary to test the program in a simulated environment before deploying it on the actual robot. For this purpose, we will utilize Gazebo [42] and RViz [43] for simulation and testing.

The Mixed Reality Toolkit 3 (MRTK3) [44] is a framework provided by Microsoft, designed to facilitates the development of XR applications within the Unity environment. It supports various XR platforms, such as Hololens 2, Meta Quest, SteamVR, etc. We will use MRTK3 to develop UI interventions for robot LfD.

3.2 Phase 0 - Performing Pick-and-Place Tasks

This phase aims to train the robot to accurately pick up an object and place it into one of the bins. It can be divided into three stages: (1) training data collection, (2) model training and (3) performance evaluation.

3.2.1 Training Data Collection

We will collect two sets of training data, each representing trajectories of moving objects towards two specific bins. The collection methodology for each set will remain consistent. Generally, there are three methods to obtain training data: kinesthetic teaching or teleoperation via human manipulation, and generation through trained agents.

Kinesthetic teaching typically yields the highest-quality training data, because the operator physically guiding the robot to execute the desired behavior in the real world. In contrast, teleoperation, especially when performed in a simulated environment, tends to introduce more noise, requiring post-processing such as smoothing. Meanwhile, it doesn't require the manipulator to be physically present with the robot, facilitating remote, large-scale data collection. Lastly, while generating data from agents might

be the most convenient method, it often struggles with complex tasks and is primarily restricted to simulated environments.

The robosuite framework (refer to section 2.2.1) offers encapsulated controllers for teleoperation, requiring either a keyboard or a SpaceMouse for input. However, to ensure optimal data for subsequent training, we plan to adopt kinesthetic teaching for data collection. We aim to produce 50 demonstrations for each training dataset.

3.2.2 Model Training

Mandlekar et al. [13] presented *robomimic*, a framework that provides a range of training datasets for five distinct tasks along with various learning algorithms. This framework also offers several pre-trained models tailored to these tasks. Notably, one of the tasks—picking up a coke can and placing it in a target bin—bears a resemblance to our task. While our research does not focus on refining learning algorithms for trajectories and the desired task is relatively straightforward, we will directly adopt the transformer architectures included in the framework. Given the distinct shapes between coke cans and cubes, we have decided to train our model from scratch rather than fine-tuning the provided pre-trained models. Two separate models will be trained, each dedicated to placing objects into one of the bins.

3.2.3 Performance Evaluation

Initially, the trained models will be assessed within a simulated environment. A variety of objects will be randomly positioned before the robotic arm. For each object, the robot will attempt to relocate it to the designated bin using the corresponding model. Failures can manifest as an inability to lift the object, move the object, or identify the target zone. Performance will be gauged by calculating the success rate. If the accuracy reaches 100%, the models will be deployed to real robots for subsequent evaluation. If this benchmark is not met, additional data collection might be necessary to enhance performance. The ultimate goal at this phase is to achieve an accuracy of nearly 100% in a real-world setting.

C

3.3 Phase 1 - Learning Sorting Rules

The goal of this phase is to leverage MRC to enable the robot to learn sorting rules by observing demonstrations. It can be divided into two stages: (1) implementing object detection and (2) machine learning (ML) for behaviours.

3.3.1 Implementing Object Detection

Before teaching the robot specific sorting rules, it is essential to detect the region where the object is positioned. To implement object detection, we will first gather training data using MRC. In Hololens 2, two holographic regions will be delineated by the user: one in red and the other in blue. The goal is to identify which region the object overlaps, thus determining its category. We will utilise object detection algorithms such as YOLO (You only look once) and R-CNN, benchmarking them based on accuracy and response time metrics. It's worth noting that response time is anticipated to significantly influence the overall user experience.

3.3.2 Machine Learning for Behaviours

Objects can vary in shape and color. Their features, combined with their respective categories, will serve as training data for learning the targeted sorting behaviors. Given this task can be regarded as a binary classification problem, a range of fundamental ML techniques can be applied. Our initial choice is the ID3 (Iterative Dichotomiser 3) decision tree because of its transparent interpretation of the learning outcomes, which is presumed to be suitable for our scenario. Subsequently, we'll assess the combined efficacy of object detection and the decision tree using the F1-score. Alternative ML methods will be considered if the decision tree underperforms.

3.4 Phase 2 - Integration and Study 1

Phase 0 focuses exclusively on the robotics aspect, whereas Phase 1 is entirely dedicated to the XR dimension. The objective of Phase 2 is to integrate the robotic arm with the Hololens 2. Subsequently, Study 1 will be conducted to assess the system's overall performance. This will address RQ1 to find out if MRC can be leveraged in the robot LfD process.

3.4.1 Integrating XR with the Robot

Figure 10 illustrates the fundamental structure of the integrated system. During the demonstration, the MRC is streamed to the object detection unit to determine the corresponding region of each object. Subsequently, training data is fed into the machine learning unit to infer the user-specified sorting criteria. Upon completing this learning phase, the user can direct the robotic arm in Hololens 2 to classify the objects according to the specified rules.

The workflow of the robotic arm is depicted in figure 11. It commences by selecting the designated object and lifting it. Utilizing the previously learned criteria, it then employs the corresponding trained model to transport the object to the desired area. Finally, the robotic arm places and releases the object. This sequence iteratively continues until either all items have been sorted or the user decides to terminate the operation.

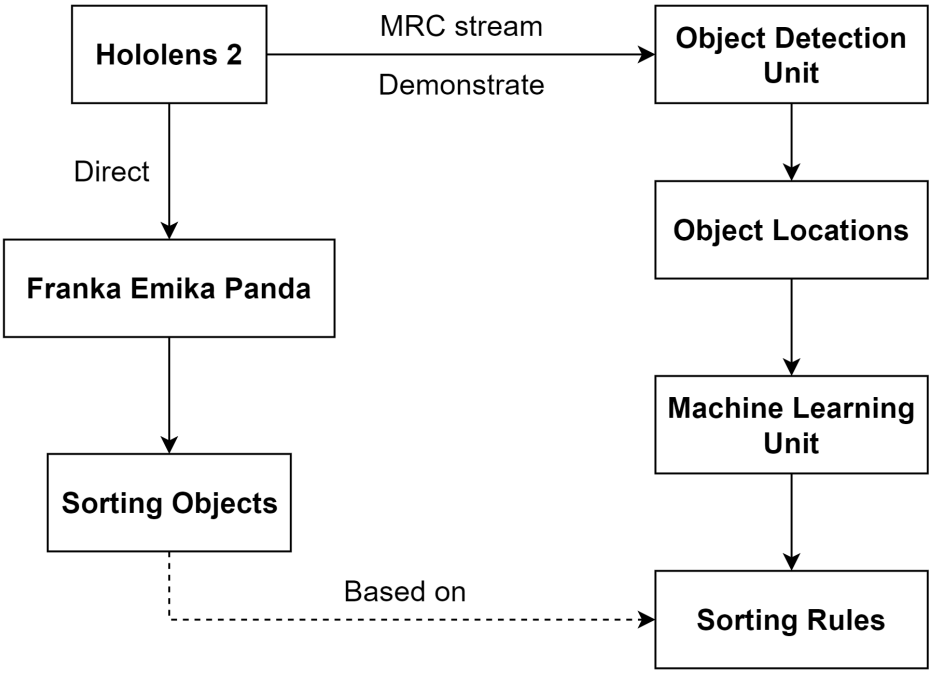

Figure 10: Fundamental system structure

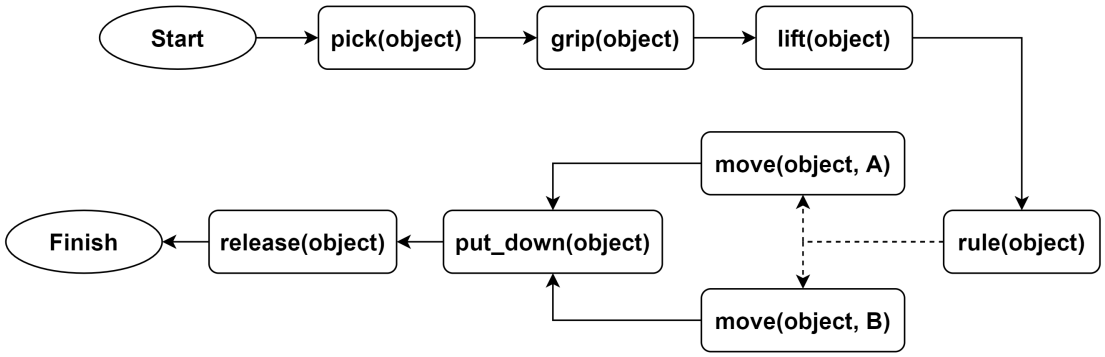

Figure 11: Workflow of the robotic arm

3.4.2 Study 1 - LfD using MRC

Study 1 is designed to address RQ1 by examining the efficacy of MRC integration in robot LfD. Multiple sorting rules will be set up for evaluation. The overall performance of the integrated system will be assessed based on two criteria: the success rate of the sorting task and the average execution time for a single workflow. This performance will be compared with benchmarks from conventional methods. If the results are satisfactory, it could prove the potential of integrating MRC into robot LfD. Such integration might pave the way for novice users to generate visual demonstrations more actively in subsequent applications.

3.5 Phase 3 - UI Interventions and Study 2

The target of this final phase is to incorporate UI interventions within the system, aiming to improve the teaching efficiency. At last, study 2 will be undertaken to evaluate the performance of these novel interventions.

3.5.1 Implementing UI Interventions

Taking inspiration from Phaijit et al. [10], the UI interventions can be designed based on four aspects derived from the formalisation mentioned in section 2.2.2. The following outlines potential UI interventions for each of these aspects.

Intervention on the demonstration set (b): The teacher may create incorrect or excessive demonstrations. To address this issue, we aim to implement two potential solutions. The first solution is to allow the teacher to modify or remove past demonstrations. This is assumed to enhance fault tolerance during demonstration creation, thereby improving teaching efficiency. Another approach is to alert the teacher if they are about to create a demonstration that closely resembles a previous one, thus preventing the generation of unnecessary demonstrations and thereby enhancing teaching efficiency. Demonstration similarity can be measured using a predefined threshold δ:

$$sim(\beta_n, \{\beta_1, ..., \beta_{n-1}\}) \leq \delta \tag{9}$$

Intervention on the teacher's bias (Q): The teacher's bias is influenced by various factors that originate from human nature. We will primarily focus on the teacher's understanding of the robot's operational abilities and their mental model. We will not take other latent factors, such as the teacher's emotional state, into consideration.

To assist novice users in becoming familiar with the robot's operational capabilities, previous studies have involved visualizing the physical constraints of the robot [3]. We intend to visualize the reach range of the robotic arm by overlaying it on the real world. An alternative option is to visualize a virtual replica of the robot in Hololens 2. This can, for example, prevent users from placing objects outside of the robot's reach range.

Assisting users in comprehending the robot's mental model serves as a means to bridge the information gap in HRI. Previous research has explored methods such as using waypoints, arrows, semantic explanations, and adding virtual arms or eyes on robots to convey the robot's intent [37, 45, 46]. Phaijit et al. [10] also conducted tests where they attempted to directly display the decision tree, but users may have had difficulty comprehending it. We plan to develop an interface that converts the ML model into natural language, showing the rules the robot learned from observations.

Intervention on the observed task realisation (R): Once a set of demonstrations is established, allowing the instructor to preview the robot's learning outcomes could guide decisions on any subsequent demonstrations that might be needed. Implementing a virtual replica of the robotic arm can facilitate the teacher in assessing the current learning status, considering that previewing on the actual robot could cost more time and resources. Furthermore, the teacher can pose specific queries to the robot regarding particular objects to verify its sorting accuracy. It is anticipated that these interfaces will assist the instructor in refining their demonstration strategies.

Intervention on teacher's interpretation of the desired behaviour ($\tilde{B}$): The teacher may have a different understanding of the desired task compared to the robot. $B \setminus \tilde{B}$ denotes the tasks overlooked by the teacher. For example, the teacher might forget to illustrate a sort rule for a specific category of objects. An intervention can be applied to notify the user in such scenarios, aiming to augment the completeness of the demonstrations. Besides, $\tilde{B} \setminus B$ represents the behaviours expected by the teacher but which fall outside the defined task space. For example, the teacher might demonstrate an object not present within the set of existing objects. Our focus will be placed on $B \setminus \tilde{B}$, as the size of $\tilde{B} \setminus B$ is limited.

3.5.2 Study 2 - Impact of UI Interventions

We plan to recruit 30 participants with no prior background in robotics to conduct a user study, investigating the impact of various UI interventions introduced. Before commencing, participants will be presented with several predefined sorting rules, though

they will also have the option to craft their own. Additionally, they will have a tutorial on the task process, fundamental device operations, and UI navigation.

Performance will be evaluated across three dimensions: (1) outcomes of the sorting task, (2) teaching efficiency, and (3) users' understanding of the learning status. Both accuracy and F1-score will be recorded to assess the sorting task results. For an in-depth discussiong on teaching efficiency (η) and user's interpretability (i_a) of the learning status, kindly refer to section 2.2.2.

$$\eta = \frac{\varepsilon}{|b|}, \eta \in [0, 1] \tag{10}$$

$$i_a = \frac{|R \cap \tilde{R} \cap B| + |B \setminus (R \cup \tilde{R})|}{|B|}, i_a \in [0, 1] \tag{11}$$

Our primary objective is to determine which intervention aspect most significantly enhances teaching efficiency. Participants will be divided into five groups: one group will operate without any interventions, while each of the remaining groups will be assigned a distinct intervention aspect. For an initial statistical evaluation comparing each intervention against the no-intervention group, we will employ the Kruskal-Wallis H-test. Subsequent post-hoc analysis will utilize the Dunn's Test. For a horizontal comparison of the interventions, we'll assess the F1-score and accuracy to explore which has the most pronounced impact.

In the final stage, we plan to investigate whether the combination of all these interventions can yield improved performance. An additional group will be designated, with all interventions enabled, to assess the previously mentioned metrics. The outcomes will provide insights into addressing RQ2, examining whether the implementation of these UI interventions in XR can enhance teaching efficiency.

4 Contribution and Significance

In this research, we introduce an innovative method in HRI that utilizes MRC to generate visual demonstrations for robot LfD. This approach is particularly user-friendly for novice users as it eliminates the need for specialized knowledge in robot manipulation. By leveraging the features of XR, the procedure of LfD can bypass physical limitations, simplify the demonstration creation process and provide users with a completely immersive experience. Notably, users are not required to be in close proximity with the robot, allowing for remote teaching. Furthermore, as demonstrations are generated from the user's viewpoint, this method holds the potential to address the correspondence problem inherent in learning behaviors from visual cues. Advanced developments in computer vision and LLMs suggest that more complex behaviours, such as cooking or assembling, could be efficiently encoded through mere observation, presenting promising avenues for subsequent research in this approach. In essence, our study aims to highlight the promise of MRC as a potent tool in robot LfD.

Another significant contribution of this research lies in our design and implementation of novel UI interventions specifically tailored for robot LfD. These interventions are designed to bridge the information gap between the human instructor and the robot, aiming to improve the overall teaching efficiency. Through the proposed user study, we expect to assess the system's performance with novice users, gaining insights that will shape future refinement of these interventions. While current exploration of XR applications in robot LfD remains to be limited, our efforts may suggest that XR could serve as a crucial role in promoting the use of robot LfD in people's daily lives. This could also potentially pave the way for a broader adoption of robots and XR headsets in commercial applications.

5 Timeline

The follow-up research timeline is scheduled as shown in Figure 12, commencing from week 8 of semester 2, 2023.

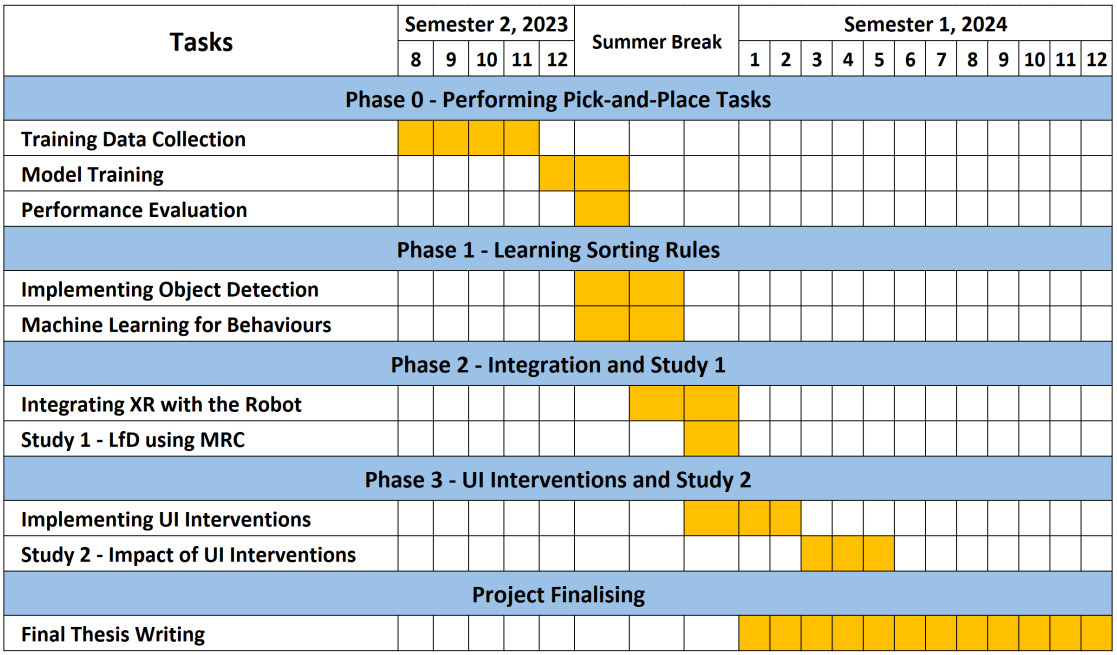

Figure 12: Timeline for the proposed research

References

[1] 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.

[2] Bradley Hayes and Brian Scassellati. Challenges in shared-environment human-robot collaboration. *learning*, 8(9), 2013.

[3] Matthew B Luebbers, Connor Brooks, Carl L Mueller, Daniel Szafir, and Bradley Hayes. Arc-lfd: Using augmented reality for interactive long-term robot skill maintenance via constrained learning from demonstration. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3794–3800. IEEE, 2021.

[4] Matthew B Luebbers, Connor Brooks, Minjae John Kim, Daniel Szafir, and Bradley Hayes. Augmented reality interface for constrained learning from demonstration. In *Proceedings of the 2nd International Workshop on Virtual, Augmented and Mixed Reality for HRI (VAM-HRI)*, 2019.

[5] Michael Walker, Thao Phung, Tathagata Chakraborti, Tom Williams, and Daniel Szafir. Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy. *ACM Transactions on Human-Robot Interaction*, 12(4):1–39, 2023.

[6] Maximilian Speicher, Brian D Hall, and Michael Nebeling. What is mixed reality? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[7] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–33, 2022.

[8] Stefanie Zollmann, Christof Hoppe, Tobias Langlotz, and Gerhard Reitmayr. Flyar: Augmented reality supported micro aerial vehicle navigation. *IEEE transactions on visualization and computer graphics*, 20(4):560–568, 2014.

[9] Roy D Pea. User centered system design: new perspectives on human-computer interaction. *Journal educational computing research*, 3(1):129–134, 1987.

[10] Ornnalin Phaijit, Claude Sammut, and Wafa Johal. User interface interventions for improving robot learning from demonstration. Unpublished manuscript, The University of New South Wales, Sydney, Australia, 2023.

[11] Microsoft. Mixed Reality Capture Overview, 2022.

[12] Carl Mueller, Jeff Venicx, and Bradley Hayes. Robust robot learning from demonstration and skill repair using conceptual constraints. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6029–6036. IEEE, 2018.

[13] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

[14] Paul Milgram, Herman Colquhoun, et al. A taxonomy of real and virtual world display integration. *Mixed reality: Merging real and virtual worlds*, 1(1999):1–26, 1999.

[15] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997.

[16] Ornnalin Phaijit, Mohammad Obaid, Claude Sammut, and Wafa Johal. A taxonomy of functional augmented reality for human-robot interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 294–303. IEEE, 2022.

[17] James E Young, Min Xin, and Ehud Sharlin. Robot expressionism through cartooning. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 309–316, 2007.

[18] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. Communicating robot arm motion intent through mixed reality head-mounted displays. In *Robotics Research: The 18th International Symposium ISRR*, pages 301–316. Springer, 2020.

[19] Cheng Guo, James Everett Young, and Ehud Sharlin. Touch and toys: new techniques for interaction with a remote group of robots. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 491–500, 2009.

[20] Jared A Frank, Matthew Moorhead, and Vikram Kapila. Realizing mixed-reality environments with tablets for intuitive human-robot collaboration for object manipulation tasks. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 302–307. IEEE, 2016.

[21] Hooman Hedayati, Michael Walker, and Daniel Szafir. Improving collocated robot teleoperation with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 78–86, 2018.

[22] Tomonori Yamamoto, Niki Abolhassani, Sung Jung, Allison M Okamura, and Timothy N Judkins. Augmented reality and haptic interfaces for robot-assisted surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 8(1):45–56, 2012.

[23] Dennis Krupke, Frank Steinicke, Paul Lubos, Yannick Jonetzko, Michael Görner, and Jianwei Zhang. Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018.

[24] Microsoft. Eye tracking on HoloLens 2, 2022.

[25] Atsushi Watanabe, Tetsushi Ikeda, Yoichi Morales, Kazuhiko Shinozawa, Takahiro Miyashita, and Norihiro Hagita. Communicating robotic navigational intentions. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5763–5769. IEEE, 2015.

[26] Microsoft. Voice input, 2022.

[27] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.

[28] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[29] Sylvain Calinon, Florent Guenter, and Aude Billard. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(2):286–298, 2007.

[30] 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.

[31] Jacopo Aleotti and Stefano Caselli. Part-based robot grasp planning from human demonstration. In *2011 IEEE International Conference on Robotics and Automation*, pages 4554–4560. IEEE, 2011.

[32] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

[33] Rüdiger Dillmann. Teaching and learning of robot tasks via observation of human performance. *Robotics and Autonomous Systems*, 47(2-3):109–116, 2004.

[34] Erik A Billing and Thomas Hellström. A formalism for learning from demonstration. *Paladyn, Journal of Behavioral Robotics*, 1(1):1–13, 2010.

[35] Rolf Pfeifer and Christian Scheier. *Understanding intelligence*. MIT press, 2001.

[36] Aran Sena and Matthew Howard. Quantifying teaching behavior in robot learning from demonstration. *The International Journal of Robotics Research*, 39(1):54–72, 2020.

[37] Maximilian Diehl, Alexander Plopski, Hirokazu Kato, and Karinne Ramirez-Amaro. Augmented reality interface to verify robot learning. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 378–383. IEEE, 2020.

[38] Franka Emika. Product Manual Franka Emika Robot, 2021.

[39] Microsoft. Microsoft HoloLens, 2023.

[40] Franka Emika. Franka Control Interface, 2017.

[41] ROS. ROS 2 Documentation: Humble, 2023.

[42] Gazebo. Gazebo Homepage.

[43] ROS. RViz - ROS Wiki, 2018.

[44] Microsoft. Mixed Reality Toolkit 3, 2023.

[45] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafir. Communicating robot motion intent with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 316–324, 2018.

[46] Jared Hamilton, Thao Phung, Nhan Tran, and Tom Williams. What's the point? tradeoffs between effectiveness and social perception when using mixed reality to enhance gesturally limited robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 177–186, 2021.