Augmented Reality Human-Robot Interface for Assisting Robotic Manipulation

A Major Qualifying Project

submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

degree of Bachelor of Science

By:
Dimitri Saliba
Tyler Giroux
Alexander Sun
Bryon Tom
Justin Kyi

Date: May 2025
Report Submitted to:

Professor Jane Li, Advisor
Professor Koksal Mus, Advisor
Worcester Polytechnic Institute

Augmented Reality Human-Robot Interface for Assisting Robotic Manipulation

Dimitri Saliba
ECE/RBE
dasaliba@wpi.edu

Tyler Giroux
RBE
tjgiroux@wpi.edu

Justin Kyi
RBE
jkyi@wpi.edu

Bryon Tom
CS/RBE
brtom@wpi.edu

Alexander Sun
CS/RBE
acsun@wpi.edu

Abstract—**Robot Autonomy has imprecision in both perception and action. Human assistance can correct this inaccuracy by teaching a robot to make corrective adjustments. The Microsoft HoloLens 2 is an Augmented Reality headset that allows users to view and manipulate holograms in three dimensions, which is well suited for intuitive Human Robot Interaction. However, its tracking capabilities are poorly tailored for fine robot control. Our team proposes a human robot collaborative workspace, as well as a collection of AR Human Robot Interfaces designed to facilitate human assisted robotic manipulation. In addition, we designed a compound manipulation and perception task showcasing our proposed platform to assist robot autonomy and propose a user study to evaluate its efficacy.**

Index Terms—**digital twin, augmented reality, mixed reality, microsoft hololens 2**

I. INTRODUCTION

The precise operation of robotic arms can be a strenuous task for human operators. Even under the best conditions, issues with how a robot perceives its environment are also prevalent, with operators experiencing issues with regard to depth perception and misjudging distances. These issues are exacerbated by a lack of intuitive communication between the system and the operator.

The rising prevalence of Augmented Readlity (AR) technology provides a unique solution to the problem of unintuitive human-robot interaction. Through the usage of visual elements, the robot is able to intuitively communicate its perception of the workspace to its user, enabling better interactions between the robot and its operator. Recent applications have proven that the visual information provided by the AR interface improves task performance with regard to robotic manipulation, reducing physical demand on the operator while simultaneously reducing task completion time. Existing AR interfaces struggle to colocate between the AR device, the robot, and the workspace, with these solutions struggling with precision in static workspaces where typical SLAM strategies don't work.

Our project aims to create a set of multimodal AR interfaces through the HoloLens tailored for human assisted robotic manipulation. Through the use of a multi-camera vision pipeline, the robot will obtain a more accurate idea of where objects are within its workspace, allowing the user to direct the robot in a controlled manner.

II. RELATED WORK

In the last few years, there has been a surge in research utilizing Virtual Reality (VR), AR, and Mixed Reality (MR). One major factor in this is the growing popularity of commercial head-mounted VR, AR, and MR (VAM) devices such as the HoloLens, Oculus Quest, and Apple Vision Pro. This development of cheaper, and more accessible systems, along with more extensive libraries and support for development, means that VAM visualization holds great promise for the field of Human Robot Interaction (HRI). Numerous studies have shown that VAM can illustrate a wide range of robots' navigation and perception abilities. For example, AR can display a robot's navigation status, destination, and intended path, facilitating communications between a robot and its user [2]. This addresses one of the more significant limitations of robot operation, that being the difficulty of achieving two-way communication between a robot system and its user. Recent studies have only supported this concept, with results concluding that AR-assisted robot operation possessed reduced difficulty compared to its non-assisted counterpart while decreasing the time to completion of the task.

While some robotic systems can operate autonomously, many require human assistance for updating, calibration, and error correction. These actions can be time-consuming and are only as good as certain diagnostics can provide. Communication between humans and robots is essential for these issues, which is something HRI can drastically improve. Effective communication between humans and robots should be bilateral, where both sides provide additional information to the alternate party. Augmented human perception of the robot is the information given to the human from the robot through AR. A common example of this would be a robot displaying and marking obstacles in its planned trajectory. This data would be otherwise unknown to a user but can be easily displayed with expanded communication options [3]. Similarly, augmented robot perception of the human is information given to the robot from the human. There are many examples of this information transfer, with most revolving around other methods of human communication such as speech, gaze, and gestures. For example, a robot could infer a human's goal based on gaze detection of the object the human is looking at and could be further expanded by awaiting verbal confirmation before executing an action. [3]

Recent studies, such as ones that were documented by Zhanat Makhataeva and Huseyin Atakan Varol, have been observing the numerous fields where VAM has been applied. In their research, they classified four major categories from the five years of 2015 to 2019. The categories they established were:

1) Medical robotics: Robot-Assisted surgery (RAS), prosthetics, rehabilitation, and training systems
2) Motion planning and control: trajectory generation, robot programming, simulation, and manipulation
3) Human-robot interaction (HRI): teleoperation, collaborative interfaces, wearable robots, haptic interfaces, brain-computer interfaces (BCIs), and gaming
4) Multi-agent systems: use of visual feedback to remotely control drones, robot swarms, and robots with shared workspace [4]

These fields saw large improvements in metrics in their respective categories, such as accuracy, safety, consistency, expanded study, and improved control. While there were a lot of noted improvements, Makhataeva and Varol also made note of the shortcomings of the current VAM hardware and software. They highlight the main areas where they think improvement is necessary, finding that current wearable devices have limited field-of-view, poor tracking stability, and crude user interfaces [4]. With these presented issues, potential future research routes were proposed, including a wider field of view and improved resolution, more advanced interfaces, and improved methods of object localization and registration. Our group's project aimed to tackle the last two points of future research by incorporating a system of co-location that utilizes spatial anchors and fiducial tags. In addition, multiple user interfaces were designed and implemented in a compound task to allow for increased communication between the user and the robot.

Spatial anchors are very commonly used in both augmented and mixed reality. They are essential for spatial mapping and collaborative visualization but are not as precise as needed for certain tasks. The accuracy of spatial anchors is heavily dependent on the precision and reliability of the sensors used for tracking. A dynamic workspace is also challenging when working with spatial anchors, as moving objects or obstacles can disrupt stability as the environment changes. In a paper written by Wennan He, Mingze Xi, Henry Gardner, Ben Swift, and Matt Adcock, they discuss how a study fused spatial anchors with fiducial tags to create a consistent and effective tracking system for assets in a warehouse [5]. They utilized spatial anchors to track the approximate location of assets, and when needed, utilized the fiducial tags to provide a more precise location. This fusion of tracking methods provides a way to cancel out some of the noisy readings obtained exclusively using spatial anchors. Our project aimed to incorporate a similar system for dynamically tracking objects in a collaborative workspace. Due to the many changes our workspace underwent, our group found that spatial anchors alone were insufficient for our needs. After incorporating a similar system to the one evaluated by He, Xi, Gardner, Swift, and Adcock, our group saw significant improvement in tracking capabilities.

III. SYSTEM

In this section, we describe the system that includes the workspace and the Augmented Reality interfaces.

A. *Compound Task*

The objective of the system is to complete a compound multimodal task, which requires human-robot collaboration. The task begins with the user placing the blue and white cups on the table in the workspace of the robot. The system will detect the blue cup sitting stationary and prompt the user through the HoloLens' audio if the robot should pick it up. Upon receiving verbal confirmation from the user, the robot will hover its end-effector above the handle of the cup and then ask the to make the final adjustment for pickup using an AR interface through the HoloLens. Once the user specifies the final position of the end-effector, the robot will pickup the cup and place it above the white cup. Then the robot will display the one-shot adjustment interface in the HoloLens. This interface will require the user to specify the true pour location of the blue cup to the robot, upon which the robot will make the specified adjustment. Following that movement, the user must gaze at the robot and say the ”pour” verbal command to trigger the dial interference to appear. The dial is then used to rotate the wrist of the robot and precisely pour the contents from one cup to another. Once the pouring is complete, the user will move the box into the workspace and the HoloLens will prompt the user via voice command to end the task by placing the blue cup into the box.

The task centers around the pouring sub-task due to its natural complexity for a robot to complete on its own. The alignment of the two cups for pouring is often an action that's difficult for a robot to infer and with various mediums in the cup may act sporadically upon pouring without accounting for visual or haptic feedback. Thus the robot requires precise human intervention to complete the task which ensures that the interfaces are robust and intuitive.

B. *Workspace*

Fig. 1 shows the workspace of our system with areas of interest expanded.

The first area of interest is the **Trossen Robotics Locobot** with 5 DOF arm highlighted in aqua located in the bottom left corner. Our team chose to utilize the robot due to time and budgetary constraints on the project. Through the pyrobot libraries much of the basic manipulation mathematics and

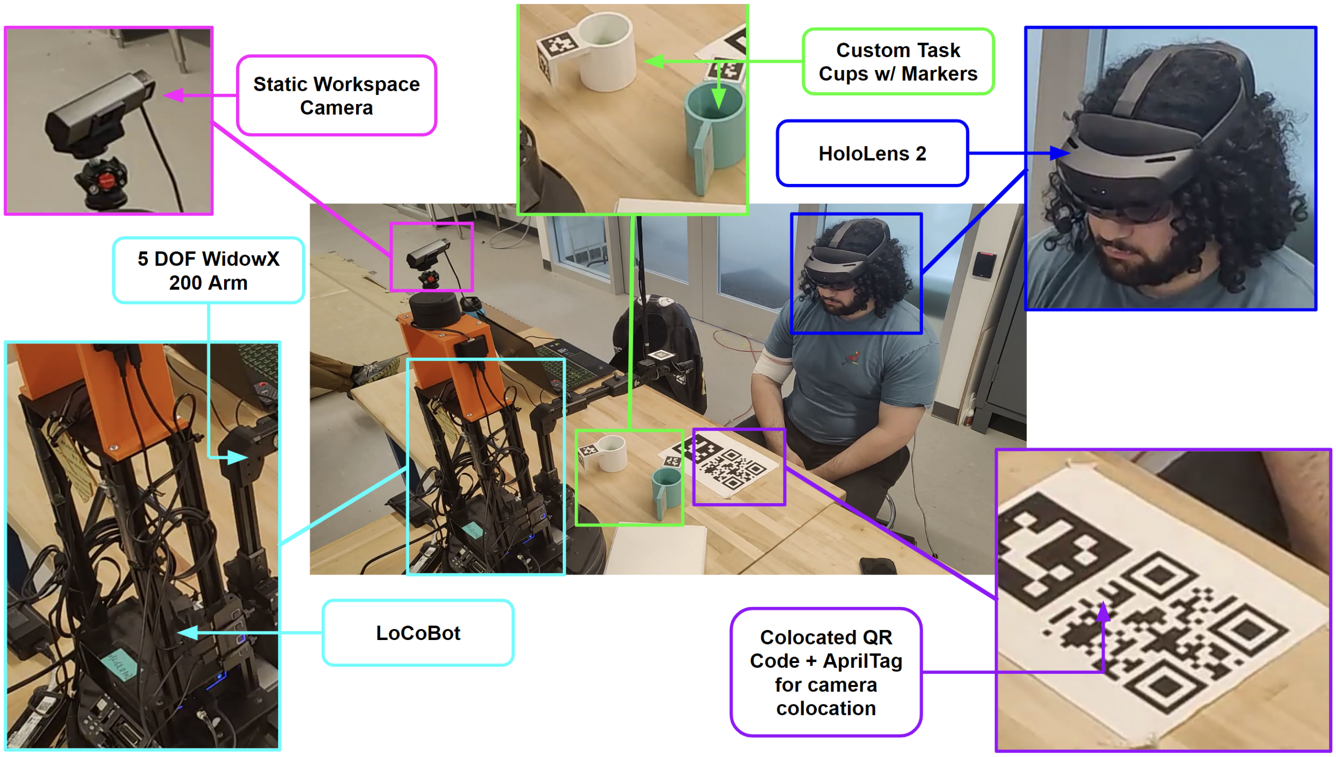

Fig. 1. Control flow of Yes or No interface.

functionality was already implemented which allowed our team to rapidly improve the other portions of the system.

In the top right corner in blue, a user is wearing the **Microsoft HoloLens 2** headset used to command the system via Augmented Reality interfaces.

Our system utilizes two **Static Workspace Cameras**, one of which is shown in pink in the top left corner. One camera has a 2k resolution and runs at 60 fps, whilst the second (not shown) has a 4k resolution and runs at 30 fps. These two cameras provide fused continuous pose estimates for all tabletop objects marked with AprilTags.

Two of the main tabletop objects that the robot interacts with are the **3D printed cups** that are marked with fiducial markers. They are shown in the top middle of the figure and are marked in green. These cups comprise the pickup and pouring portions of the later compound task and display the need for a precise perception system and robust human-robot interfaces.

The final aspect of Figure 1 is the **Colocation Tags** that are shown in purple and located in the bottom right of the image. This piece of paper may seem rather odd at first, however, the AprilTag acts as an origin so that the system will calibrate each camera's position relative to this point. This allows the position estimates for each object coming from each camera to be synced to the same coordinate frame. The second image on the paper is a QRCode which is used by the HoloLens to place its Spatial Anchor at a precise point, with a known fixed transformed to the origin of the workspace. The HoloLens currently only supports QRCode marker detection for static objects which works well for the placement of the anchor.

The system also includes a box (not shown) that is marked with an April Tag, which is used to store the blue cup and signals the end of the task. These different pieces of the workspace all play a crucial part in executing the collaborative workspace environment between human and robot.

C. Software Architecture

The diagram in Fig. 2 shows a high level overview of our software organization. First, we have the workspace ROS shown in the green box where the object detection occurs. This set of ROS nodes contains the camera pipeline where our two workspace cameras read all of the objects in the workspace are marked with April Tags. Each camera creates its own estimate for each marked object using OpenCV and publishes these estimates to our ROS TF tree used to help visualize and confirm object and camera estimates. The separate estimates are then fused into a set of final position estimates in an intermediate publisher which also creates estimates of certain components of our objects such as the centers and handles. This all gets processed, packaged, and passed on to the Unity Engine or HoloLens via our Unity publisher. The Unity Engine then aligns it's own coordinate frame with the ROS workspace using a Spatial Anchor and our colocation tag mentioned in the workspace section.

The Unity Engine powers all of the HoloLens functions necessary for our task including rendering holographic overlays

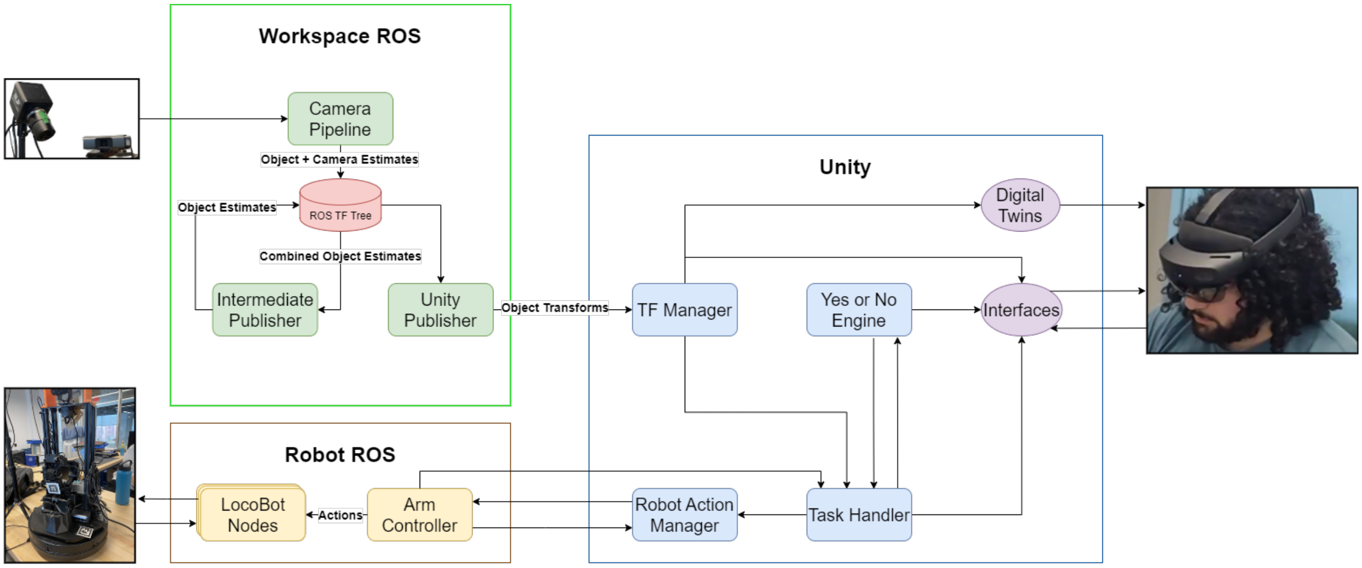

Fig. 2. Overall software control flow showing how Unity communicates with all ROS nodes

of objects as well as handling the various interfaces which are gaze, voice command, and gesture controls. Additionally, the Unity Engine contains a state machine for our task which is in charge of commanding the Robot to perform manipulation tasks as well as controlling which interfaces are used at different stages of the task.

Finally, we have the Robot ROS workspace which takes the commands given by the Unity Engine in the form of ROS messages to perform manipulation tasks. Since the Robot ROS only takes in ROS messages, it is in charge of communicating specific arm commands to the built-in LocoBot nodes. For example, during the initial pickup phase the HoloLens only prompts the user if the cup should be picked up while the robot performs most of the pickup portion autonomously.

D. Interfaces

Augmented Reality Objects: One of the most compelling uses of AR for Human Robot Collaboration is the ability to easily convey the robot's perception of the environment. Fig. 3 illustrates how Holographic Objects are used in our system to help the Human better visualize the Robot's understanding of the workspace. For example, the object twins shown in blue convey to the Human which task objects the Robot can currently see, as well as the quality of its localization. Frame axes are also drawn along the robot joints to illustrate the quality of the colocation between the Robot and the HoloLens.

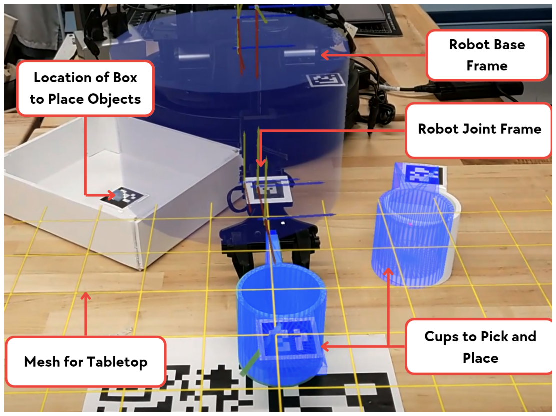

Fig. 3. User's mixed reality view.

Yes or No Question: Inferring user intent is a common problem in Human Robot Collaboration, and false negatives can lead to a frustrating user experience, while false positives can lead to unwanted and potentially dangerous unexpected behavior. Our proposed solution to this issue is an interface, which utilizes Yes or No questions to validate user intent. The use of speech recognition as opposed to a gesture-based confirmation menu was chosen to reduce physical fatigue and more accurately emulate Human-to-Human collaboration. For scalability, we abstracted this system, creating the schema shown in Fig. 4. The question schema includes a prompt to be read to the user, as well as functions to invoke depending on the answer. This allows for easy creation and alteration of questions. By default, questions will timeout if left unanswered for 10 seconds, and upon timing out questions will invoke their "onNo" function.

Robot Oriented Control: Another common use for Human Assistance in robotic manipulation is error correction. There are many reasons why an autonomous pick action may fail, such as inaccuracy in object tracking, inaccuracy in collocation, or other factors such as unknown displacement from an object tracker to a viable grip location. We designed an AR interface to correct these inaccuracies via gesture control. The interface shown in Fig. 5 allows real-time position control of the Robot's end effector for pickup alignment. The interface can be moved by pinching the white halo, as the interface is moved it sends its relative displacement from its starting

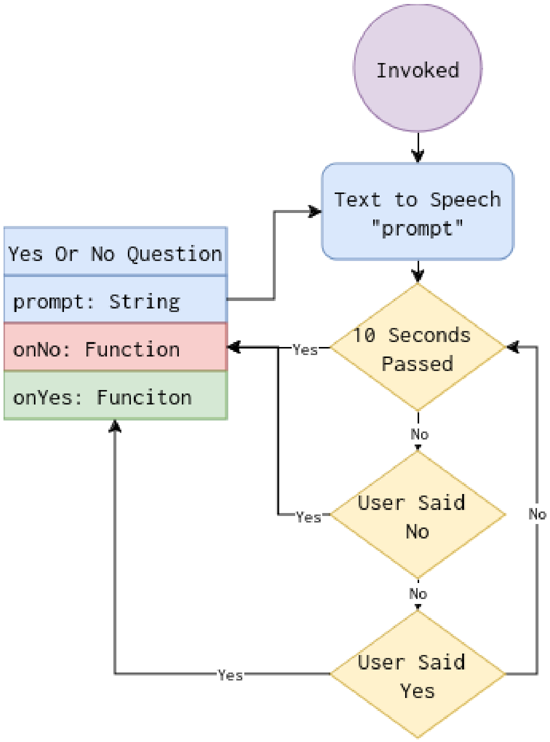

Fig. 4. Control flow of Yes or No interface.

position to the Robot. By sending locations relative to the starting position of the end effector this interface is agnostic to colocation error. Upon being released the interface signals to the Robot to continue the pick action.

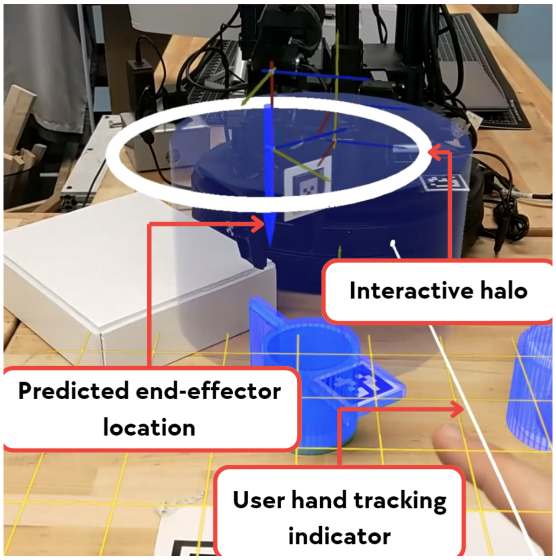

Fig. 5. Robot Oriented interface for real time position control.

Object Oriented Control: Sometimes object properties or dynamics are very simple for humans but can be complicated to model accurately for use in Robotic automation. An example of this examined in our compound task is the act of pouring objects from one cup to another. Since the trajectory of poured objects is far easier for humans to accurately estimate, we designed an interface for specifying object oriented transforms. The interface shown in Fig. 6 allows the user to specify a single transform via gesture control. By dragging the white halo the user aligns the middle beam with the location poured objects will land. The Robot can then use the inverse of this transform to adjust its end effector location.

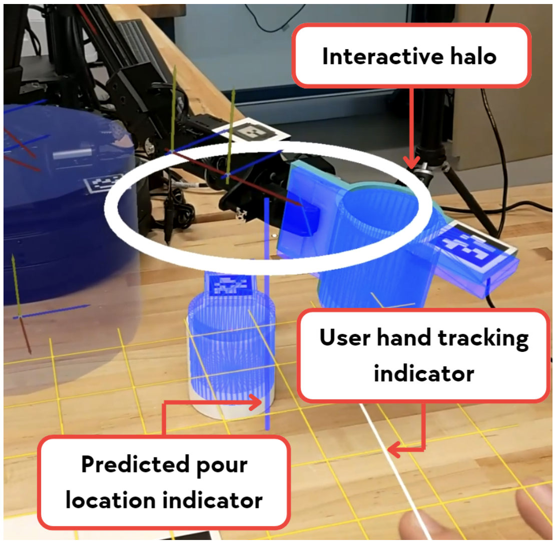

Fig. 6. Object Oriented position adjustment interface.

Additionally, for cases such as the proposed pouring task where the transform supplied by the user will remain constant relative to the object the interface can store and recall previously provided displacements. The logic shown in Fig. 7 describes how the system recalls requested adjustments if needed.

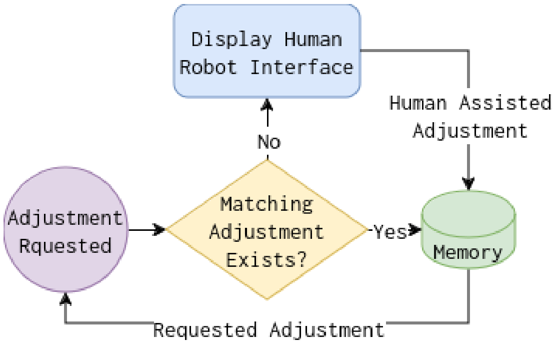

Fig. 7. Logic flow for recalling previous adjustments.

Constrained Motion Control: While the ability for 3-Dimensional input is usually a strength of AR in cases where precise movement in specific axes is necessary. For example, in our task, while pouring objects from one cup to another accidental translation or rotation could be detrimental to the success of the pour. For this reason, we designed a Radial Dial interface for real-time constrained motion control. The dial shown in Fig. 8 is comprised of a central readout which shows the current angle in degrees, along with a white sphere the user can move via gesture control to specify a rotation, however, the motion of the white sphere is constrained along the blue track, ensuring that only one rotational axis is affected. The rotation of the end effector is updated in real-time, allowing for fine grain and responsive control.

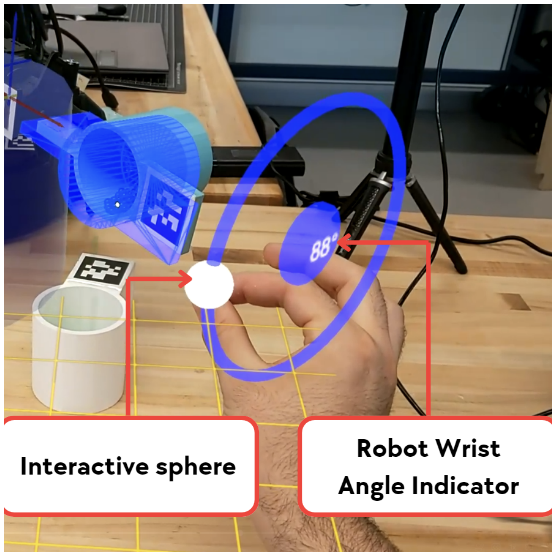

Fig. 8. Radial Dial interface for real time constrained radial motion.

Gaze Speech Fusion: Our system utilizes voice commands for initiating sub-tasks, such as the use of the command ”Pour” to display the Radial Dial interface and begin pouring. However, the use of pure speech control can be prone to false positives, which can lead to unexpected and potentially dangerous robot actions. Our team addressed this issue by fusing gaze and speech control. Speech commands issued to the robot will be ignored unless the user's gaze intersects the holographic overlay of the robot. This is done to better emulate Human-to-Human collaboration, where typically eye contact is used along with speech to indicate the recipient of a command or question without the explicit use of their name. When the user's gaze is on the robot, the overlay changes color to indicate to the user that it is ready to accept commands.

IV. DISCUSSION

While our system was able to accurately complete its task, there are many opportunities for refinement. A substantial issue we faced came with quirks of the HoloLens, namely gesture detection difficulties and spatial anchor inaccuracies. The HoloLens would have issues with hand tracking in a lot of random but specific situations which affected how we designed our gesture-based interfaces. Thoughtful interface design helped mitigate a majority of the initial issues we had in performing the tasks associated with gesture-based interfaces. Along with this, we found that during long development sessions with the HoloLens, the spatial anchor would drift creating inaccuracies within our system requiring a HoloLens reboot.

Although our interfaces performed well, we could have had a more robust combination of interfaces. For example, we used voice commands to confirm the user's intent for pickup portions. We could have also used them for the interfaces themselves in case the user encountered a scenario when an interface was not needed.

The dual workspace camera system worked well, with estimated object positions being quite accurate. Even when the robot arm obscured the object tag, the singular remaining camera was able to accurately estimate the object's position. Greater camera quality has the potential to allow for more accurate objection detection as well as a greater refresh rate for live tracking. However, there were a few inaccuracies with object detection which showed in the HoloLens when the digital twins of objects would be less accurate in some workspace areas but not others. There was also a noticeable delay in live tracking caused by delays in image acquisition which we could not entirely mitigate.

V. FUTURE WORK

We developed a system that can be built upon, with several extensions from this point on. The focus would be on expanding to more complex tasks using the robot. This could be in more accurate manipulation or by implementing movement through the mobile base, which was not pursued during this project. By extension, new interfaces would have to be implemented to demonstrate the Hololens' ability to assist the robot using digital twins. Our team has discussed the possibility of implementing trajectory visualization through holograms, allowing the user to edit the robot's suggested trajectory. This could also be done for the manipulator's trajectory as well, with the interface allowing the robot to deal with untracked objects in the workspace.

Another future work could be replacing the workspace camera system with the robot's onboard lidar or onboard camera. Colocating the HoloLens to a SLAM based mapping from the robot could pose a very interesting problem to solve with the benefits of no longer having a limited workspace or needing the current calibration routine. This could also provide a new interesting interface that facilitates this colocation process.

Finally, running a user study would allow us to measure the efficacy of our current system and interfaces. A possible user study could compare performing our task with the HoloLens vs. complete manual control (i.e. teleoperation via keyboard or game controller). There would be an administrator that runs participants on how the different interfaces work and allow participants a 5-10 training time for each system to get familiar with them. Then the participant would run through the task with both systems while the administrator collects data based on the following evaluation metrics:

1) Task completion time
2) Percentage of BBs poured into the white cup
3) Accuracy of picking up the blue cup

4) Situational awareness tested by the administrator asking questions during the task

Upon completion of the experiment, the participant will complete the NASA Task Load Index questionnaire to evaluate the mental workload of each system along with additional feedback regarding intuitiveness and preferences between systems. This would be a significant step for the project due to gathering important metrics for various aspects of the system. This feedback could be used to better design digital twins, interfaces, and robot routines, resulting in more efficient communication between both the human and robot parties.

VI. CONCLUSION

Human-robot collaborative environments can provide robots with a finer level of control and precision than purely autonomous systems. They can provide methods that can correct for foundation level faults in both action and perception. The platform that our team proposes, along with various AR interfaces through the Hololens 2, allows users to intuitively supervise, direct, and control the robotic manipulation of a compound pick and place task. The robot's own perception and choice of actions is imperfect on it's own, but with the created interfaces through gaze detection, voice commands, and digital twins, the robot can be assisted to complete the tasks with high fidelity, while possessing both a low learning curve and a lesser cognitive workload for the user.

Our team utilized the Microsoft HoloLens 2, Spatial Anchors, and April Tags to sync digital twins with their real world counterparts. Through utilizing our proposed method of co-location and new user interfaces, users can more easily assist robots in completing delicate tasks.

REFERENCES

[1] Walker, Michael, 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*, no. 4 (2023): 1-39.

[2] Suzuki, Ryo, 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*, pp. 1-33. 2022.

[3] Phaijit, Ornnalin, 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)*, pp. 294-303. IEEE, 2022.

[4] Makhataeva, Zhanat, and Huseyin Atakan Varol. "Augmented reality for robotics: A review." *Robotics 9, no. 2* (2020): 21.

[5] He, Wennan, Mingze Xi, Henry Gardner, Ben Swift, and Matt Adcock. "Spatial anchor based indoor asset tracking." *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (2021): 255-259.

[6] Ostanin, Mikhail, Stanislav Mikhel, Alexey Evlampiev, Valeria Skvortsova, and Alexandr Klimchik. "Human-robot interaction for robotic manipulator programming in Mixed Reality." In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2805-2811. IEEE, 2020.

[7] Chan, Wesley P., Geoffrey Hanks, Maram Sakr, Haomiao Zhang, Tiger Zuo, HF Machiel Van der Loos, and Elizabeth Croft. "Design and evaluation of an augmented reality head-mounted display interface for human robot teams collaborating in physically shared manufacturing tasks." *ACM Transactions on Human-Robot Interaction (THRI)* 11, no. 3 (2022): 1-19.

[8] Chan, Wesley P., Camilo Perez Quintero, Matthew KXJ Pan, Maram Sakr, HF Machiel Van der Loos, and Elizabeth Croft. "A multimodal system using augmented reality, gestures, and tactile feedback for robot trajectory programming and execution." *Virtual Reality (2022)*: 142-158.

[9] Chan, Wesley P., Geoffrey Hanks, Maram Sakr, Tiger Zuo, HF Machiel Van der Loos, and Elizabeth Croft. "An augmented reality human-robot physical collaboration interface design for shared, large-scale, labour-intensive manufacturing tasks." In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11308-11313. IEEE, 2020.

[10] Arevalo Arboleda, Stephanie, Franziska Rücker, Tim Dierks, and Jens Gerken. "Assisting manipulation and grasping in robot teleoperation with augmented reality visual cues." In Proceedings of the *2021 CHI conference on human factors in computing system*s, pp. 1-14. 2021.

[11] Ong, Sean, Varun Kumar Siddaraju, Sean Ong, and Varun Kumar Siddaraju. "Azure Spatial Anchors." *Beginning Windows Mixed Reality Programming: For HoloLens and Mixed Reality Headsets (2021)*: 175-188.

[12] He, Wennan, Mingze Xi, Henry Gardner, Ben Swift, and Matt Adcock. "Spatial anchor based indoor asset tracking." *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (2021): 255-259.

[13] Angelopoulos, Angelos, Austin Hale, Husam Shaik, Akshay Paruchuri, Ken Liu, Randal Tuggle, and Daniel Szafir. "Drone brush: Mixed reality drone path planning." In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 678-682. IEEE, 2022.

[14] Marvel, Jeremy A., Shelly Bagchi, Megan Zimmerman, and Brian Antonishek. "Towards effective interface designs for collaborative HRI in manufacturing: Metrics and measures." *ACM Transactions on Human-Robot Interaction (THRI) 9, no. 4 (2020)*: 1-55.

RESOURCES

Here are the relevant resources for anyone interested in, expanding upon or continuing the project:

- ROS Github - https://github.com/technoAl/ROS-Nursing-Robot-MQP-23-24
- Unity Github - https://github.com/dsaliba/hololens_unity_workspace
- Locobot Github - https://github.com/technoAl/Locobot-MQP/tree/master
- Google Drive - https://drive.google.com/drive/folders/1kxjzlJixN_vR490iaCVHHyYqUpfUIbbW?usp=sharing