

# HoloSpot: Intuitive Object Manipulation via Mixed Reality Drag-and-Drop

Pablo Soler<sup>1,\*</sup>, Petar Lukovic<sup>1,\*</sup>, Lucie Reynaud<sup>1</sup>, Andrea Sgobbi<sup>1</sup>, Federica Bruni<sup>1</sup>, Martin Brun<sup>1</sup>, Marc Zünd<sup>1</sup>, Riccardo Bollati<sup>1</sup>, Marc Pollefeys<sup>1,2</sup>, Hermann Blum<sup>1,3,†</sup> and Zuria Bauer<sup>1,†</sup>

<sup>1</sup>ETH Zurich <sup>2</sup>Microsoft <sup>3</sup>Lamarr Institute / Uni Bonn

{psoler, plukovic, lreynaud, asgobbi, fbruni, mabrun, mazuend, rbollati  
pomarc, blumh, zbauher} @ethz.ch

**Abstract**— Human-robot interaction through mixed reality (MR) technologies enables novel, intuitive interfaces to control robots in remote operations. Such interfaces facilitate operations in hazardous environments, where human presence is risky, yet human oversight remains crucial. Potential environments include disaster response scenarios and areas with high radiation or toxic chemicals. In this paper we present an interface system projecting a 3D representation of a scanned room as a scaled-down ‘dollhouse’ hologram, allowing users to select and manipulate objects using a straightforward drag-and-drop interface. We then translate these drag-and-drop user commands into real-time robot actions based on the recent Spot-Compose framework. The Unity-based application provides an interactive tutorial and a user-friendly experience, ensuring ease of use. Through comprehensive end-to-end testing, we validate the system’s capability in executing pick-and-place tasks and a complementary user study affirms the interface’s intuitive controls. Our findings highlight the advantages of this interface in improving user experience and operational efficiency. This work lays the groundwork for a robust framework that advances the potential for seamless human-robot collaboration in diverse applications. Paper website: <https://holospot.github.io/>.

## I. INTRODUCTION

Human-robot interaction (HRI) is becoming more and more important due to increasing deployment of robots and the complementary nature of human and machine capabilities. Robots are now commonly deployed in scenarios where human presence is risky or impractical, such as hazardous environments and remote locations. Examples include industrial settings with toxic materials, areas affected by natural disasters, and space exploration missions. These deployments minimize the need for on-site human intervention, enhancing operational efficiency and safety [26], [6].

A significant challenge in HRI lies in the interfaces used to control robots. Traditionally, three-dimensional (3D) environments are represented on two-dimensional (2D) screens. This representation can complicate the operator’s ability to maintain situational awareness and precision by limiting its depth perception and field of view. This makes it especially challenging to interact with and manipulate objects accurately in a 3D space [2]. Mixed reality (MR) technology offers a promising solution by providing a more immersive

\*Shared first authorship. †Equal supervision. This project was partially funded by the ETH Foundation. We thank Jiaqi Chen for providing guidance to develop the HoloLens app and all our participants for the user study.



Fig. 1: **Visualization of Our System.** Left: A user wearing a Hololens to interact with a holographic representation of the environment. Right: Robot pick-and-place sequence of actions. Top: Optimal grasp selection with AnyGrasp.

and intuitive interface for human-robot interaction. This approach addresses the limitations of traditional 2D interfaces by providing a 3D visual representation that aligns with the real-world environment for better scene understanding.

Recent research has explored the use of MR headsets to control robots. For example, Chen et al. demonstrate real-time visualization and intuitive navigation of robots in 3D, highlighting the benefits of MR interfaces over traditional methods [7]. These studies underscore the advantages of MR technology in enhancing the interaction between humans and robots, making it possible to perform more complex and precise tasks.

Building on this, our project presents an innovative design for a user-friendly MR application, enabling users to command a robot to perform pick-and-place tasks. The primary goal of our project is to create an interface that allows users to control a mobile object manipulator through natural, intuitive actions in a MR environment.

Our system achieves this by implementing several key features. First, we scan and reconstruct the deployment environment in 3D. Second, we segment the reconstruction into individual object instances through 3D object instance segmentation [24], [27]. Third, we make the digital twin interactive based on the object segmentation and link any interaction in the digital twin to a robot that reproduces the same environment interaction in the real environment. Therefore, by simply dragging and dropping virtual objects, users can instruct the robot to pick up and move these objects

to specified locations in the real world. To ensure ease of use, our application includes an interactive tutorial that guides users through the interface and functionality. This tutorial helps users quickly get familiar with the system, making it accessible even to those with minimal technical expertise.

The main contributions of this work are:

- We present a novel interface for remote robot operations that enables object manipulation using a drag-and-drop feature within a MR environment.
- We implement the proposed interface as an app for the HoloLens and the Spot, featuring an interactive tutorial, movable objects, and operable drawers.
- We perform extensive real-world tests with our interface and a user study, demonstrating that our interface is intuitive and enables users to perform complex pick-and-place tasks with minimal training.

## II. RELATED WORKS

The intersection of Mixed Reality and Human-Robot Interaction has become a dynamic research area, with studies exploring the potential of MR to enhance robotic control and human collaboration in complex environments. We highlight a selection of recent and particularly pertinent studies.

### A. Control via Mixed Reality Interfaces

Human-Robot Interaction (HRI) has been transformed by recent advances in Mixed Reality (MR) technologies, especially in intuitive control and situational awareness. Early work, such as [7], showcased how 3D visualization in MR enhances real-time robotic system control, bridging the gap between abstract mechanisms and direct manipulation. This intuitive spatial understanding set the foundation for later developments. Alternatively, [17] delved into robotic navigation with MRNaB, leveraging MR beacons to improve human-robot cooperation in intricate spatial settings.

In parallel, [25] categorized AR's influence on HRI, with particular attention to its advantages in enhancing teleoperation and task management in complex environments. This contributes to ongoing MR interface research that seeks to enhance precision and fluidity in interaction. [13] pushed these developments further by incorporating digital twins and haptic feedback, adding tactile input to the visual interface. This system improves control precision, especially in scenarios where visual data alone is insufficient. Their work represents a significant evolution in MR, highlighting its potential to create more effective HRI systems.

### B. 3D Instance Segmentation and Grasp Pose Estimation

One of the main challenges in robotics lies in understanding and interacting with complex environments. This issue is often addressed via 3D instance segmentation and grasp pose estimation. The Spot-Compose framework [19] tackles this by using point cloud segmentation algorithms to identify objects, making it highly effective for tasks like object retrieval and manipulation in dynamic environments. In parallel, OpenMask3D [27] enhances 3D segmentation by employing open-vocabulary models. This enables robots to

interact with previously unknown objects based on natural language descriptions. This flexibility is essential for improving adaptability in real-world applications.

Simultaneously, grasp pose estimation has seen substantial progress through the development of AnyGrasp [14]. AnyGrasp predicts two-finger grasps directly from 3D point clouds, using a dense supervision strategy that incorporates real perception data and analytic labels. This method accounts for factors such as the object's center-of-mass and environmental constraints, ensuring stable and collision-free grasps. Grasp pose estimation has progressed significantly, with frameworks like GraspNet-1Billion [15] which predicts stable, collision-free grasps from 3D point clouds by analyzing object geometry and environmental constraints. This approach complements the capabilities of Spot-Compose and OpenMask3D in segmentation tasks. Moreover, [9] explored techniques based on deep learning to enhance grasping in robotic manipulation, improving the handling of objects under various physical conditions.

### C. Enhancing Human-Robot Collaboration through MR

Human-robot collaboration is critical in environments requiring precision, situational awareness, and real-time decision-making. MR interfaces have revolutionized this collaboration by enabling seamless interaction between human operators and robotic systems. Research by [12] underscores MR's versatility in improving remote control of robotic systems, particularly in environments where human presence is limited but precision is important. This approach enables operators to manage tasks more effectively, enhancing both performance and safety.

Augmented reality (AR) interfaces, as investigated by [21], have shown promise in improving situation awareness in human-robot teaming for exploring tasks. Their study evaluates how AR visualizations can support better decision-making during time-sensitive operations, questioning about the effectiveness of these tools in collaborative settings. Furthermore, [3] explored the utility of MR interfaces in contamination-critical environments. It shows how such systems enhance teleoperation accuracy and reduce error rates.

Similarly, [8] explored how MR can optimize human-robot interaction (HRI) in remote manipulation tasks, significantly lowering cognitive load while improving task precision. Collaborative MR workspaces, like those developed by [26], allow multiple users to control and monitor robotic systems simultaneously. This shared interface improves coordination and task efficiency, particularly in team-based operations requiring collaborative control.

Finally, [22] developed a system where autonomous robots detect changes in the environment and communicate these through augmented reality, enhancing human situational awareness. Their work highlights the potential of AR in supporting human-robot teams by making environmental data more accessible and understandable.

## III. METHOD

In this section, we present the main components and techniques employed by the system. An overview is illustrated

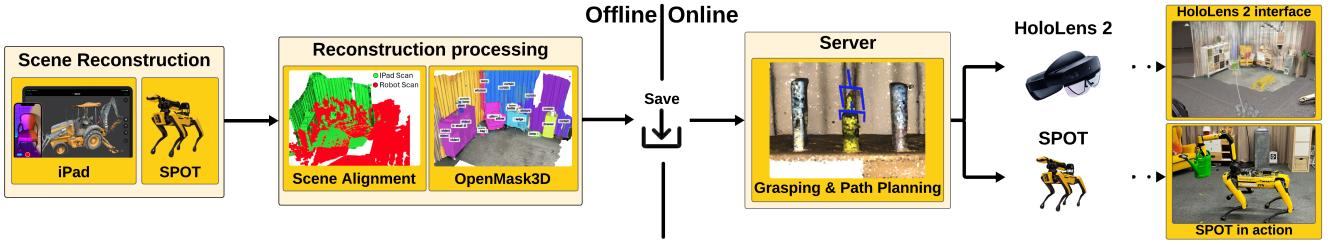


Fig. 2: **System overview.** Our method relies on both offline and online segments. Offline segment (left), is used to construct 3D scene that will be utilised in the subsequent online phase. The online segment (right) is used to control the Spot using the HoloLens with the help of an online server.

in Figure 2. In the following we go through each part of our proposed method.

#### A. Obtaining 3D Scene Reconstruction

We assume an available pre-scanned 3D representation of the scene. This representation can be obtained through SLAM with a suitably equipped robot or handheld scanners. In our experiments, we use a LiDAR-equipped iPad and the *3D Scanner App* [1] (high detail scan). Additionally, like in [19], we instruct the robot to perform a depth scan (low detail scan) with its onboard cameras. While the robot’s cameras face downward and therefore cannot be used alone to reconstruct the scene, the two registered point clouds then form a comprehensive 3D reconstruction of the scene, which is used in all subsequent steps.

#### B. 3D Semantic Instance Segmentation

We segment the 3D reconstruction into semantic instances using the OpenMask3D [27] framework. This is illustrated in subsequent steps in Figures 2 and 4. This state-of-the-art method supports open vocabulary queries on 3D scenes, allowing us to predict and delineate object instances within a 3D point cloud together with their semantic descriptors.

The result of this step is presented at the end of the Figure 4. The point cloud is segmented into each movable object instance and the static environment. It is worth noting that we provide a predefined list of language prompts to select movable objects, or segmented point cloud for that matter. All other parts of the scene are considered static.

#### C. Unity App Interface on the HoloLens

To enable intuitive interaction with the reconstructed 3D environment, we develop a Unity application (see Figure 3) for the Microsoft HoloLens 2 [20]. The interface shows the 3D reconstruction as a hologram in front of the user, where they can then with their hands drag-and-drop the segmented movable objects through the HoloLens hand tracking.

The application imports the point cloud of the scene, with each movable object instance as an independent grabbable hologram. To facilitate user manipulation, we further add a virtual floor to the hologram. When an object hologram is grabbed and subsequently dropped, custom scripts send the object’s new coordinates to the main workstation. To ensure operational feasibility, a control mechanism verifies whether the drop location is within the robot’s operational area.

During testing, we found it beneficial to include a visual representation of the robot with a status display to keep users informed of the robot’s activities. For users unfamiliar with

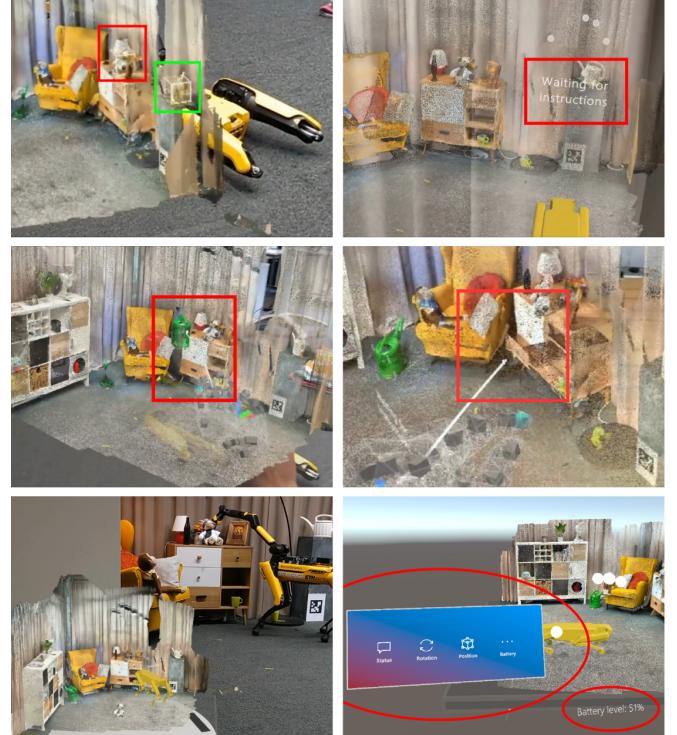


Fig. 3: **HoloLens Interface.** The above figures show various aspects of the visual interface on HoloLens. Labeling images with number left to right, and top to bottom we have: on the first image bounding boxes displaying movable objects following “show items” voice command, the following image shows how robot status is displayed during manipulation. Images four and five display user manipulating watering can and a drawer. On the fifth and sixth images we can see virtual representation of the robot and menu containing battery percentage and additional status information.

the HoloLens, we introduce a tutorial to explain the app’s features and practice drag-and-drop. We also add confirmation buttons to avoid sending wrong object placements to the robot and to alert users if the chosen location is not feasible. Additionally, we integrate voice control features: the “*show items*” command highlights movable objects to the user with colored bounding boxes, while the “*reset*” command allows users to return all moved objects to their original positions. All features are illustrated in Figure 3.

#### D. Online operation of the Robot

After an object is selected by the user and dropped at a new location in the interface, the interaction gets translated into a

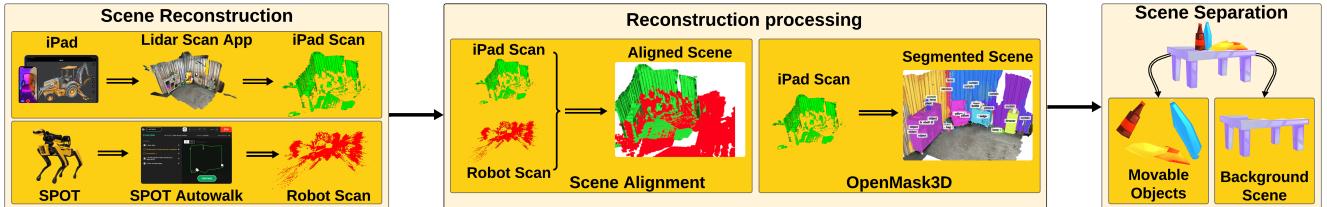


Fig. 4: **Offline pipeline.** The preprocessing done before deploying our system can be separated into three parts: Scene Reconstruction, Reconstruction Processing and Scene Separation. Scene reconstruction consists of gathering high and low resolution scans using the iPad LiDAR and the Spot cameras, respectively. Recorded point clouds are then aligned into the same coordinate system and the high resolution scan is segmented using OpenMask3D [27]. At the end we manually separate segmented instances into drappable objects and static environment.

pick-and-place command and sent to the robot. In Figure 5, this is depicted as “*Navigation & Grasping Instruction*”. The command includes the drop coordinates in the 3D scene and the index of the object in the scene representation that is available to both the robot and the interface.

To execute the pick-and-place command, we use the Spot-Compose [19] framework that we briefly outline here: Picking and placing starts with grasp estimation on the object using AnyGrasp [14]. We run inference over multiple rotations of the instance mask, since AnyGrasp identifies poses based on the frontal view. The system then performs joint optimization of poses and grasps based on the AnyGrasp [14] score, the alignment of the robot body with the grasp pose and the vicinity of obstacles. Once the robot has moved to the pose from which it should grasp the object, a local point cloud of the object is captured using the depth camera located in the robot gripper. This local capture is aligned to the initial scan using ICP [4] to obtain a corrective transformation for the grasp pose, compensating any misalignment or drift between robot odometry and scene representation (see “*Optimise grasp*” in Figure 5). Next, the robot grasps the item, picks it up, and subsequently moves to the drop location. The route to this location is calculated at the same time as the route from start to the object location, both routes are calculated using RRT [18]. We also run joint optimisation for drop location (see “*Optimise grasp*” in Figure 5). This is computed in parallel during the pick operation, since the drop location is known a priori. Finally, Spot drops the object and moves back to the start location and localizes itself again to alleviate drift.

#### E. System Integration

Our system is based on a centralized architecture, with all devices connected via WiFi (these include the Agile Mobile Robot Spot [11], the Microsoft HoloLens 2 [20] and the server). The server is responsible for the entire planning procedure and sends commands to the Spot robot using the Boston Dynamics Python SDK [10]. The AnyGrasp [14] and OpenMask3D [27] models also run on this machine. The HoloLens app queries the robot status and issues commands via a REST API exposed by the workstation.

#### F. Drawer Drag-and-Drop

To showcase the versatility of the proposed interface, we implement interactable drawers that can be opened and closed using the same drag-and-drop mechanism. Their

movement is constrained to match the range of motion of real-world drawers. This is combined with the Spot Compose functionality to allow the robot to open drawers. An example of the user interacting with the drawer is displayed in the fourth image of Figure 3.

## IV. EXPERIMENTS

To evaluate our system, we first evaluate the end-to-end reactivity and reliability of our proposed interface. This is to validate that the interface does not introduce critical additional points of failure to the mission deployment of the mobile manipulator. We then conduct a user study to compare the effectiveness of our proposed interface with a much more common screen + mouse interface.

All the experiments were performed using the Microsoft Hololens 2 [20] for the MR interface and the Agile Mobile Robot Spot [11]. For computing the grasping and path planning we used a workstation with an Nvidia RTX 4090. The commands are then sent to a device that is mounted on the robot and communicates with the Spot via ROS [23]. All communications are done with a router using static IPs.

#### A. Real World System Test

We run 240 pick-and-place actions using a variety of objects of different difficulties: Two differently sized watering cans, a mug, two toy plushies and differently shaped and colored bottles. In Table I, the different pick up locations are listed. During these trials, the robot is instructed through our HoloLens interface to pick up a specified object. Each object is moved 40 times from its original position to a fixed dropping place. For a more detailed evaluation, we divide each task in phases, which are shown in Figure 6. The two main metrics used for the real world tests are: Success rate and task duration. Success rate is responsible for measuring reliability of our system, whereas, task duration gives us a way to measure user experience and optimize our system.

**Task Duration analysis:** According to the data shown in Figure 6, the total average time of end-to-end task execution is 80 s. 17 s are taken for the path and grasp planning, which is the first step after a command is issued and therefore the most critical delay between user input and robot action. The user interaction through the drag-and-drop interface typically takes about 10 seconds, showing the effectiveness and user-friendliness of our application.

**General pick-up success:** As shown in Figure 7, on 240 trials, 172 (72%) were successful. This is overall in line

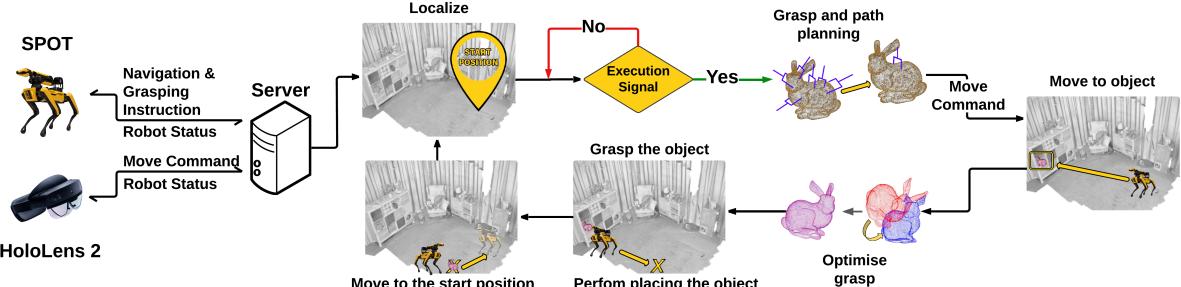


Fig. 5: **Online pipeline.** When the system is deployed (online) it follows the given loop. Start of the pick-and-place procedure is triggered by the HoloLens user when he places an object in the scene. This sends a signal to the intermediate server with additional information about the object and its location. After a successful information exchange, the robot is localized. Next, the grasp and path are calculated on the server which then sends the commands to the robot. After the robot arrived to the location, grasp optimisation is performed using ICP algorithm [4]. At the end the robot performs the grasp, moves the object and returns to the starting position where it localizes itself waiting for another trigger signal.

Object	Pick-up Location	Success Rate	Main Issue
White Can	Shelf	75%	Collision on navigation
Green Can	Floor	85%	Collision on navigation
Green Mug	Floor	70%	Empty Grasp
Black Bottle	Cabinet	80%	Object not found
Blue Plush	Cabinet	62%	Collision during grasp
Cow Plush	Cabinet	55%	Collision during grasp

TABLE I: **Object Manipulation Overview.** The different objects help to identify critical failure modes in the underlying mobile manipulation system.

with the findings from the mobile manipulation system that we use [19]. Notably, from 68 failure cases, 9 (13%) can be attributed to the user interface due to inaccurate pick-up or drop-off commands. The other failures come from the grasping (47) and navigation (12) systems, which are independent of our proposed interface. We conclude that the proposed interface does not drastically increase failure modes of the overall robotic system.

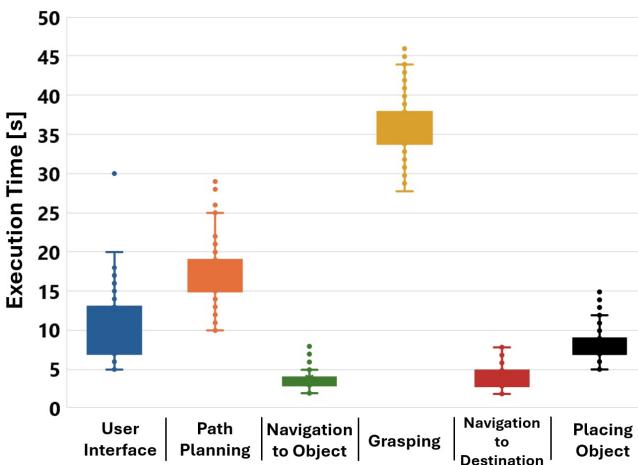
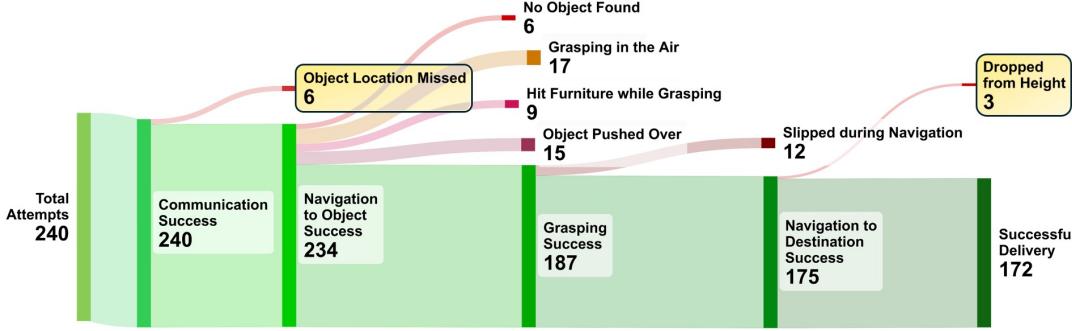


Fig. 6: **Task Completion Time Breakdown.** Mean execution times for the six stages of object placing: user interface, path planning, navigation to object, grasping, navigation to destination and placing object, respectively. Data from 240 trials; error bars show  $\pm 1\text{SD}$  and  $\pm 3\text{SD}$ .

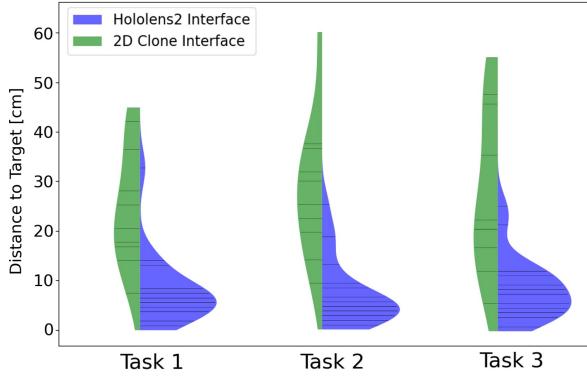
**Success rate per object:** We use 6 kind of objects for testing the different difficulties. As shown in Table I, the green watering can was the most reliably grabbed object and the cow plush animal has the lowest success rate of 55%. Different objects caused different failure modes. The green watering can, for example, collided on navigation and destination since the robot doesn't take the carried object into account for local collision avoidance. The mug is specially small, causing many empty grasps. Finally, for the stuffed cow, the main issue was due to collision during grasp. This is likely because the cow was placed in the most cluttered pick-up location. We conclude that there systematic failures for individual objects in relation to our proposed interface. It is likely that more advanced mobile manipulation systems could overcome most observed failure cases.

### B. User Study

We designed a user study to assess the intuitiveness and usability of our interface. The study was performed on a varied group of 24 participants, aging 20 to 66 years old, with technical and non-technical backgrounds. None of them had used the Hololens before. The study has been approved by the ETH Ethics Commission under project number 268. For this study, a simulated environment is used to maintain controlled experimental conditions. In the first part of the study, the participants are asked to complete the app's tutorial and are given time to familiarize themselves with the controls. In the second part, the users have to perform a set of three tasks. These tasks consisted of moving specific objects to different locations marked by visual indicators (e.g., 'Place the object on top of this shelf where you see the cross'). The third task is to command the robot to open a drawer and place a cup inside the opened drawer. Task completion rate and user accuracy are recorded for each task. User accuracy is measured as distance between predetermined goal coordinates and the actual coordinates sent by users to the robot. Furthermore, the time for completing the tutorial and finishing the tests is recorded. Finally, after the tutorial and after completing the tests, the participants are asked to respond to questions from both the NASA Task Load Index (TLX) [16] and the System Usability Scale (SUS)



**Fig. 7: Manipulation Experiments Results.** To assess the object manipulation reliability of our interface, we conducted 240 trials using six distinct objects, differing in grasping difficulty and pick-up location. The overall success rate is 72%. The most critical part is the object grasping and the primary failure is missing the object and grasping in the air. Failures highlighted in yellow are directly linked to our interface.



**Fig. 8: User Accuracy Distribution Comparison.** Comparison of the participants' accuracy at placing the objects for the three tasks. Users with our proposed HoloLens interface (blue) have significantly lower error while placing objects through drag-and-drop than users who use the 2D clone with mouse and keyboard (green). Horizontal lines mark the values for specific participants.

questionnaire [5], including questions on frustration, mental demand and ease-of-use.

To compare the performance of the proposed HoloLens setup to a traditional 2D interface for mouse and keyboard, a subsection of these users (10) also tested a 2D clone of the HoloLens app using the Unity app simulator. These participants were asked to perform the same three tasks of the HoloLens study on a monitor using mouse and keyboard. Duration, success rate, user accuracy, and the NASA TLX questionnaire answers were also recorded on this test.

The main results are shown in Table II. The average

Measure	HoloLens		Mouse & Keyboard	
	Mean	Std Dev	Mean	Std Dev
Tutorial	03:07	1:35	-	-
	4.43	1.91	-	-
	3.71	1.52	-	-
Tasks 1-3	05:56	02:02	04:47	1:42
	8.29	7.27	27.1	13.3
	8.57	6.97	31.6	15.29
	10.14	6.54	31.2	17.89
	3.76	2.43	6.5	1.77
	4.43	1.83	6.7	1.42
All	6.90	1.84	-	-
	3.57	0.93	-	-

**TABLE II: User Study Results.** Tutorial, tests, and overall measures. Comparison between the proposed HoloLens interface and the 2D clone that uses mouse and keyboard.

participant finished the tutorial in 2 to 4 minutes. It served as a short introduction to our proposed interface and to the HoloLens in general. New users do not feel overwhelmed while completing the tutorial, on average the perceived effort to finish the introductory tasks is a 3.7. Furthermore, as the participants get more familiar with the controls, their frustration levels decrease from 4.4 after the tutorial to 3.8 at the end of the study. Participants in the end felt comfortable with the interface, rating their own performance in the upper range of the scale (6.9). These results all indicate the high usability of our app, even for users who are new to the HoloLens or to virtual reality.

When comparing these results with the participants who also tested the 2D app clone, the object misplacement of the 2D counterpart is more than 3 times as high. Figure 8 visualizes these results in more detail. The majority of participants using the HoloLens are able to maintain the distance to the target under 10cm, while the 2D app clone results span from 10cm to half a meter. The main cause for the error is the user not being able to correctly estimate the depth of the object's location in the monitor. This is also results in participants reporting being more frustrated with the 2D app clone interface. The interface comparison underlines the effectiveness of our proposed interface and shows that augmented reality platforms are better suited for placement of objects in 3 dimensional spaces.

## V. CONCLUSION AND FUTURE WORK

We introduced a novel interface that enables intuitive pick and place commands for a Spot robot through a drag-and-drop in a 3D MR environment. Our solution offers an interactive and user-friendly app including a tutorial. The conducted user study shows that users without experience with MR devices can successfully operate the robot and are 3 times more accurate in object placing compared to traditional mouse and keyboard interfaces. Our experiments show that the investigated overall system allows operators to command pick-and-place tasks on a variety of objects and in different environments with a total success rate of 72%.

Possible future improvements include: Refining the drag-and-drop mechanics based on the object type, enabling the system to perform more sophisticated tasks like turning on switches, and replacing the manual list of movable object categories with a better prior e.g. from a LLM.

## REFERENCES

- [1] 3D Scanner App. <https://3dscannerapp.com/>.
- [2] Carl Ahlberg and Sara Eriksson. Challenges and developments in human-robot interaction interfaces: Overcoming 2d limitations for effective 3d control. *Current Robotics Reports*, 8(3):215–232, 2022.
- [3] B. et al. Bejczy. Mixed reality interface for improving mobile manipulator teleoperation in contamination critical applications. *Procedia Manufacturing*, 51:620–626, 2020.
- [4] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.
- [5] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.
- [6] CERN ATS. Mixed reality human-robot interface for remote operations in accelerator facilities. <https://ats.web.cern.ch/node/145>, 2023. Accessed: 2024-07-21.
- [7] Jiagi Chen, Boyang Sun, Marc Pollefeys, and Hermann Blum. A 3d mixed reality interface for human-robot teaming, 2023.
- [8] J. et al. Delmerico. Spatial computing and intuitive interaction: Bringing mixed reality and robotics together. *IEEE Robotics Automation Magazine*, 29(1):45–57, 2022.
- [9] H. et al. Duan. Robotics dexterous grasping: The methods based on point cloud and deep learning. *Frontiers in Neurorobotics*, 15, 2021.
- [10] Boston Dynamics. Spot SDK Documentation. <https://dev.bostondynamics.com/>.
- [11] Boston Dynamics. Spot: The agile mobile robot. <https://bostondynamics.com/products/spot/>.
- [12] O. et al. Erat. Drone-augmented human vision: Exocentric control for drones exploring hidden areas. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1437–1446, 2018.
- [13] Wen Fan, Xiaoqing Guo, Enyang Feng, Jialin Lin, Yuanyi Wang, Jianming Liang, Martin Garrad, Jonathan Rossiter, Zhengyou Zhang, Nathan Lepora, Lei Wei, and Dandan Zhang. Digital twin-driven mixed reality framework for immersive teleoperation with haptic rendering. *IEEE Robotics and Automation Letters*, PP:1–8, 12 2023.
- [14] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhui Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains, 2023.
- [15] H.-S. et al. Fang. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11441–11450, 2020.
- [16] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.
- [17] Eduardo Iglesius, Masato Kobayashi, Yuki Uranishi, and Haruo Take-mura. Mrnab: Mixed reality-based robot navigation interface using optical-see-through mr-beacon, 2024.
- [18] Steven M. LaValle. Rapidly-exploring random trees : a new tool for path planning. *The annual research report*, 1998.
- [19] Oliver Lemke, Zuria Bauer, René Zurbrügg, Marc Pollefeys, Francis Engelmann, and Hermann Blum. Spot-compose: A framework for open-vocabulary object retrieval and drawer manipulation in point clouds, 2024.
- [20] Microsoft. Hololens. <https://www.microsoft.com/en-us/hololens>.
- [21] T. et al. Mott. Augmented reality for enhanced situation awareness in human-robot collaborative exploration. In *4th International Workshop on Virtual, Augmented, and Mixed Reality for HRI*, 2021.
- [22] C. et al. Reardon. Enabling situational awareness via augmented reality of autonomous robot-based environmental change detection. In *VAMR 2020, Held as Part of the 22nd HCI International Conference*, pages 611–628, 2020.
- [23] Clearpath Robotics. Spot ros. <https://www.clearpathrobotics.com/assets/guides/melodic/spot-ros/>, 2020.
- [24] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation, 2023.
- [25] 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 *CHI Conference on Human Factors in Computing Systems, CHI '22*. ACM, Apr. 2022.
- [26] Krzysztof Adam Szczerk, Raul Marin, Eloise Matheson, Jose Rodriguez-Nogueira, and Mario Di Castro. Multimodal multi-user mixed reality human–robot interface for remote operations in hazardous environments. *IEEE Access*, 11:17305–17333, 2023.
- [27] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation, 2023.