



**Universiteit
Utrecht**

Multimodal Immersive Systems for Assembly in Mixed Reality

Luca Becheanu - 7630379

Supervisors: Dr. Wolfgang Hürst, Dr. Julian Frommel

Department of Information and Computing Sciences
Game and Media Technology
Master Thesis

July 3, 2024

Contents

Abstract	4
Acknowledgments	5
1 Introduction	6
2 Related Work	8
2.1 Extended Reality in Assembly	8
2.2 Devices Used During Assembly	10
2.3 Multimodal Artificial Intelligence	10
2.4 Evaluation Metrics	12
2.4.1 Technical Aspects and Tool Features	12
2.4.2 Levels of Acceptance and Participant Attitude	13
2.4.3 Psychological Reactions	13
2.4.4 Level of Immersion	13
3 Research Question	14
4 Methodology	15
5 Design and Architecture	15
5.1 Use Case Analysis	15
5.2 Requirements	16
5.2.1 User Requirements	17
5.2.2 Technical Requirements	17
5.3 System Architecture	18
6 Development	27
6.1 Feasibility Study	27
6.1.1 Technical Feasibility	27
6.1.2 Operational Feasibility	27
6.1.3 Time Feasibility	28
6.1.4 Summary of Feasibility	28
6.2 Risk Analysis	29
6.3 Implementation	29
6.3.1 Technical Details	30
6.3.2 Application Walkthrough	36
7 User Study	37
7.1 Variables	37
7.2 Hypotheses	37
7.3 Experiment Setup	37
7.4 Tasks	39
7.5 Subjects	40
7.6 Results	41
7.6.1 Pre-training results	41
7.6.2 Quantitative results	43
7.6.3 Post-training results	47

8	Discussion	50
8.1	Subquestions	50
8.1.1	How can we design an architecture that will take into account the advantages of dialogue agents and MR for assembly?	50
8.1.2	How do we integrate dialogue agents with interactive MR systems such that they are compatible and still highly performant?	50
8.1.3	How can we evaluate the efficacy of AI-based MR training?	51
8.2	Shortcomings and Future Work	51
9	Conclusion	53
	References	54
A	Appendix	61
A.1	Application Enlarged Figures	61
A.2	Experiment	67
A.2.1	Informed Consent Form and Demographics Questionnaire	67
A.2.2	Tool Usability Form	70
A.2.3	Task Load Index Form	74
A.2.4	Preferred Instruction Medium Form	75
A.3	Results	77

Abstract

This paper explores the integration of Extended Reality and multimodal Artificial Intelligence (AI) in the context of assembly tasks. The study investigates the use of dialogue agents within Mixed Reality (MR) environments to enhance assembly processes, focusing on whether training can be simplified and costs can be reduced by replacing the human trainer. Methodologically, the research includes a use case analysis, feasibility study, and system implementation, followed by a user study evaluating variables such as task performance and user satisfaction. Results indicate that integrating dialogue agents with MR systems may potentially improve assembly efficiency and user interaction, although challenges in technical development and user acceptance remain. The proposed method provides a good starting point for understanding the potential of AI-driven MR applications in training and operational settings, suggesting a course of action for further research and development.

Acknowledgments

I would like to express my deepest gratitude to my parents for their unwavering support and encouragement throughout these years. Their belief in me has been a constant source of motivation.

I extend my heartfelt thanks to Dr. Wolfgang Hürst for his invaluable help and guidance throughout this project. I am also deeply grateful to him for providing access to university hardware, which was essential for the research.

My sincere appreciation goes to the Centrum Wiskunde & Informatica (CWI) for their initial guidance, which helped shape the direction of this work, as well as Dr. Julian Frommel for his supervision.

1 Introduction

The assembly phase is a pivotal component of the overall fabrication process for two main reasons. Firstly, it represents a substantial portion of manufacturing costs, accounting for an average of 30-40% of the total expenditure [4]. Secondly, the efficiency of assembly tasks affects the final quality of the product, overall production time, and cost. Despite the notable automation in processes like cutting, milling, and forming, assembly remains predominantly conducted manually [62].

To help boost worker productivity, Extended Reality (XR) applications for supporting task execution during manufacturing have been a subject of academic research for many decades, and such tools have been built in Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), yet there are still limited examples of their actual implementation in industries [14, 51]. The most recent studies highlight the need for virtual assistant tools that provide the trainees with a means of asking for information [43], the same way as they would do in the current industry standard of face-to-face training [82].

Building these types of assembly helper software requires being aware of the specific set of closely related instructions that produce the final result [11]. The variability of product configurations in terms of components is very large, which forces workers to be highly adaptive in performing tasks that could vary slightly but significantly in terms of steps and operations [7]. With this state-of-the-art action recognition [24, 75, 87], next-generation artificial intelligence (AI) assistants are capable of handling multimodal inputs (e.g., vision, history of previous interactions, and the user’s utterances), and performing multimodal actions (e.g., displaying a visual guide while generating the system’s utterance) [44]. Combining these models that can make the dialogue agent understand more than just one type of information, such as visuals and text in a shared embedding space could ultimately provide assembly trainees with the best solution to resolving ambiguity and correcting mistakes during assembly, enhancing accuracy in task completion [34, 63].

The research aims to address this gap by constructing an AI-assisted immersive system for assembly training and determining whether it is more effective than traditional methods such as face-to-face training, introducing novel prospects in the domain of personal assistant application development, with the general goal of replacing human-to-human assembly traineeship with human-to-machine automation. We will first create an application that will take into account the advantages of superimposing virtual objects onto the physical world. Then, we will build up the interaction between the application and the dialogue agent capable of generating human-like instructions. Last but not least, we will investigate whether these systems can substitute human trainers in assembly tasks, thereby reducing the substantial resources dedicated to worker training, and evaluate the resulting improvements in work efficiency. Specifically, this study seeks to tackle the following three research questions:

- How can we design an architecture that will take into account the advantages of dialogue agents and MR for assembly?
- How do we integrate dialogue agents with interactive MR systems such that they are compatible and still highly performant?
- How can we evaluate the efficacy of AI-based MR training?

The upcoming sections are structured in the following way: Section 2 provides a review of the background literature and Section 3 presents the research questions related to the gaps found in the previous work. Following this, Section 4 describes the methodology used to tackle the problem, in Section 5 we discuss the design and architecture of the system, while Section 6 focuses on development and implementation. Section 7 reports the user study and the experiment results, in Section 8 we analyze to what extent the research questions have been answered and propose areas for future exploration, and finally, Section 9 summarizes our main contributions.

2 Related Work

This chapter is a comprehensive overview of existing research in XR, devices used during assembly, recent developments in multimodal AI, and the evaluation metrics associated with these technologies. Section 2.1 delves into the integration of XR technologies in assembly processes. It examines how AR, VR, and MR are employed to enhance efficiency and accuracy in assembly tasks. Studies highlight the impact of XR on training, task guidance, and overall assembly performance. Section 2.2 focuses on the hardware aspect, reviewing the various devices utilized in assembly processes. It covers modern technologies as a new toolset for workers, such as phones and tablets, smart glasses, and HMDs (head-mounted displays). The literature surveyed discusses the advantages and limitations of these devices, their impact on ergonomics, and the overall effect on assembly line productivity. Section 2.3 introduces Multimodal AI, exploring the synergies between different sensory modalities (e.g., vision and speech) and their potential for integration into assembly tasks. The research investigates how AI algorithms leverage multimodal inputs to enhance decision-making and adaptability in dynamic assembly environments. Section 2.4 is dedicated to the methodologies and metrics used to assess the success and impact of XR in assembly tasks. It reviews both quantitative and qualitative evaluation metrics, such as task completion time, error rates, user satisfaction, and learning curves. The literature highlights the importance of selecting appropriate metrics to measure the specific goals of the implemented technologies and offers insights into the challenges of evaluating complex, interconnected systems.

2.1 Extended Reality in Assembly

XR is defined as "a unifying concept to interpolate between the realities and to eXtrapolate beyond them" [36], considered an umbrella term coined in 1991 to refer to AR, VR, and MR. Table 1 gives an overview of the key differences of each XR technology.

	Virtual reality (VR)	Augmented reality (AR)	Mixed/merged reality (MR)
Display device	Special headset or smart glasses	Headsets optional	Headsets optional
Image source	Computer graphics or real images produced by a computer	Combination of computer-generated images and real-life objects	Combination of computer-generated images and real-life objects
Environment	Fully digital	Both virtual and real-life objects are seamlessly blended	Both virtual and real-life objects are seamlessly blended
Perspective	Virtual objects will change their position and size according to the user's perspective in the virtual world	Virtual objects behave based on user's perspective in the real world	Virtual objects behave based on user's perspective in the real world
Presence	Feeling of being transported somewhere else with no sense of the real world	Feeling of still being in the real world, but with new elements and objects superimposed	Feeling of still being in the real world, but with new elements and objects superimposed
Awareness	Perfectly rendered virtual objects can't be distinguished from the real deal	Virtual objects can be identified based on their nature and behaviour, such as floating text that follows a user	Perfectly rendered virtual objects can't be distinguished from the real deal

Table 1. Key differences between VR, AR, and MR as defined by McMillan et al. [39].

VR creates its own artificial three-dimensional (3D) environment in which the user wearing a headset is immersed, having a physical presence in the virtual world [71]. This technology thrives in training simulations of specific environments that can't be replicated in real life such as calamities [16], or that are very expensive to repetitively create in the real world such as extinguishing fires in firefighter training [18].

AR is a technology that superimposes computer-generated images, sounds, or other sensory information on a user's view of the real world, allowing them to see an overlay of digital information on physical world elements while keeping the real world environment central [12]. Unlike VR, AR adds projections and holograms to the existing world as it is. Many studies have been done to showcase its capability to give guidance and assist workers during manufacturing tasks [14, 79].

MR is the latest technology that combines aspects of both the real-world and digital elements, merging them to produce an enriched interactive environment such as simulating a virtual interactable workplace in the physical world [72]. While some recent MR papers researching assembly training [8, 57] still refer to the Reality–Virtuality continuum proposed by Milgram and Kishino [42] in 1994, a newly revised definition was put forward by Rauschnabel et al. [55] which argues that MR lies on the AR continuum where users experiences can range from a very low functional level (Assisted Reality) to highly interactive and realistic experiences (MR). The user usually wears an HMD and can interact with physical and virtual items at the same time. This technology also allows the user to immerse in this combination of worlds using their own hands, without the explicit need for any other controllers [14]. The 3D content projected from the headset will react to the user the same way as it would in the real world. Since this is the newest immersive technology, the use cases for assembly tasks are still under development [33, 76, 86], yet new advancements in the most recent headsets like the Apple Vision Pro [25] might pave the way forward to a new type of Reality that lets the user switch back and forth between virtual, augmented and mixed realities depending on use cases.

Manufacturing Phases	Tasks	Useful XR Technology in Training
Introductory Phase	safety training, orientation training, planning and designing of new tasks	VR, MR
Learning Phase	sorting, picking, keeping, assembling, installation	VR, AR, MR
Operational Phase	inspection, packing, monitoring assembly line, assembly	MR
Tangent Phase	using rare tool/machinery, hand tool, power tool	AR, MR
End Phase	cleaning routine (process, shovel, sweep, clean work areas), inspection	AR, MR

Table 2. XR technology usage in training as reviewed by Doolani et al. [14].

Our focus is on MR since it is the most flexible technology that can be used in all manufacturing phases, and the only technology that can be used during the operational phase of the manufacturing process that entails assembly as seen in Table 2, since it is not completely immersive like VR and lets the user interact with the real world along with digital models.

2.2 Devices Used During Assembly

In order to decide which tools fit best in the development of a system, the technological limitations of the devices used need to be taken into account. Werrlich et al. [79] survey most recent assembly studies and note that a limitation to most of these is the usage of hand-held devices such as tablets and smartphones. However, headsets and HMDs offer the trainee the possibility to work hands-free while providing users with the necessary information to perform their tasks. This influences the development of new applications on hardware that makes the trainee experience as helpful and ergonomic as possible such as HMDs, which are becoming less bulky and lighter.

Research shows that millions of employees are going to use smart HMDs on a regular basis for their on-job tasks and training by 2025 [22, 37, 59]. It is also very useful in teaching trainees to prevent the risks of injury during construction or assembly in high-consequence practical industries like healthcare [2, 45], aerospace [35], and manufacturing [32], where mistakes can be deadly [54].

Learning applications have been implemented that make use of the new possibilities of the most recent headsets [5]. Werrlich et al. [81] evaluated the efficacy of HMDs using a user study with two groups and argued that a combination of the real and virtual assembly phases strengthened the training transfer, exactly what the MR field offers. Gonzalez-Franco et al. [21] developed an MR setup and tested it in implementing an aircraft maintenance door to verify if it can replace other forms of face-to-face training. While the collaborative interaction was between two humans and not between a human and an AI model, it showed that the MR setup can potentially provide ways of collaborative training.

MR training technologies encompass more than just hardware, the effectiveness and quality of these platforms are also heavily reliant on the software they use. High-quality software is essential for operating the hardware and for generating MR content. The gap in the growth rates of software and hardware presents challenges for developers, who must continually adapt to these evolving technologies and develop applications with new functionalities [14]. Previously, applications and development engines needed separate proprietary code for each device on the market. However recently, manufacturers now have a specification they can follow to ensure their system is compatible with past, present, and future applications. Application developers no longer need to worry about target platforms as a new standard called OpenXR ensures the app will behave similarly on all conformant devices [49]. Therefore our goal is to create software in the form of a MR application that can work with any OpenXR-compliant HMD hardware.

2.3 Multimodal Artificial Intelligence

Large Language Models (LLMs) have shown significant advancements in dialogue systems and their applications [47, 67, 88]. The majority of contemporary AI systems are unimodal, meaning they are engineered to operate exclusively with a single data type and utilize algorithms specifically designed for that modality [58]. For instance, an unimodal AI system employs algorithms based on Natural Language Processing (NLP) to interpret and derive meaning from textual content. Consequently, the sole form of output this chatbot is capable of generating is text. In contrast, multimodal architectures that can integrate and process multiple modalities simultaneously have the potential to receive multiple inputs and produce more than one type of output, understanding and responding

to spoken commands while simultaneously processing visual cues from the environment.

As part of the research on existing and upcoming products, companies are currently pushing innovations towards having virtual assistants in XR that are connected to LLMs and object detection modules to help with day-to-day tasks, such as translating foreign languages, choosing the right clothing, and so on [73]. This applies to a teaching environment too, where XR is making a huge impact on eLearning and training programs, since most workers are practical learners, acquiring 70 percent of their skills and knowledge from experiential learning [30, 46, 66]. However, limited work has been done to replace the human trainer factor with AI modules. To this extent, further work is needed to optimize the process as AI technologies are evolving toward high-performant dialogue agents.

Remarkable breakthroughs in the field of computer vision have made it possible for AI models to make very specific distinctions in classifications, recognizing images belonging to a subordinate category [27, 31, 78, 85]. These new types of dialogue agents work similarly to the human visual system, which is very capable of image reasoning, telling the difference between a dog and a cat, but also distinguishing between different dog breeds, despite some being very similar to one another.

Research on AI dialogue agents in training environments is highly valuable for a future with seamless human-to-machine interaction, creating a synergy between human intelligence and AI capabilities. Making increased automation more applicable to real-world scenarios contributes to enhancing productivity and efficiency in various domains [17]. Additionally, by focusing on training applications, researchers can explore ways to enhance safety protocols, optimize production processes, and improve overall operational efficiency, thereby paving the way for a more advanced and interconnected landscape. The integration of AI dialogue agents in such settings not only streamlines tasks but also opens up new possibilities for innovation and growth in various industries [14].

Related to dialogue agents, Padmakumar et al. [50] propose task-driven embodied agents that can communicate via language. Their research provides a human-machine interaction dataset specifically tailored for household tasks, however, they focus on training the model using a human acting as the virtual agent trainer sending textual instructions to the human trainee. While this research is needed to create performant dialogue AI agents, they focus on an embodied agent in VR domains, which is an intelligent agent that interacts with the environment through a physical body within that environment [74]. However, as previously seen in Table 2, an embodied VR agent cannot be incorporated in the operational phase of assembly, due to the lack of seeing the physical parts needed to be assembled.

Xu et al. [84] made use of ChatGPT to optimize the AR-based assembly tasks and reduce the cognitive load required in analyzing complex text-based instructions, by taking pictures with the virtual camera of the headset and projecting the AI model’s instructions on a digital twin of the physical object. Currently, the most important limitation that steers our goal away from Assisted Reality applications is the lack of interaction with virtual projections. For a concrete example, even if the dialogue agent showcases the missing parts of an object on top of the real-world object, the user has no ability to physically interact with the digital parts at any given step, while participants in similar studies wanted that as an option because the application is centered around 3D interaction [43]. We would like to use the capabilities of LLMs in a similar fashion, however, the lack of an important digital touch modality is the main difference compared to what we are trying to achieve.

2.4 Evaluation Metrics

Borsci et al. [7] conducted a comprehensive review of the effectiveness of VR and MR tools for training operators, particularly in the context of car service maintenance. They identified a trend among automotive researchers to focus their analysis only on car service operators' performance in terms of time and errors. However, they noted that important pre- and post-training aspects that could affect the effectiveness of VR/MR tools to deliver training content were often left unexplored. These aspects include people skills, previous experience, cybersickness, presence and engagement, usability, and satisfaction. Their work highlights the need for a more holistic approach to evaluating the effectiveness of VR/MR tools for training, one that takes into account a wider range of factors beyond just performance metrics, as seen in Table 3.

	Evaluation criteria	References
1	Technical aspects and tool features	Bowman [9]
	Effect of designed features, expected system functioning	Stefanidis et al. [65]
2	Levels of acceptance of MR tools	Gallagher et al. [20]
	Participants attitude/engagement	Kneebone et al. [26]
		Sanchez-Vives and Slater [60]
3	Psychological reactions - cognitive load, skills, stress	Grantcharov et al. [23]
3	Psychological reactions - cognitive load, skills, stress	Witmer and Singer [83]
		Seymour et al. [61]
4	Level of immersion	Robert S. Kennedy and Lilienthal [56]
		Stanney et al. [64]

Table 3. List of evaluation criteria reported in the literature as important for testing the effectiveness of VR/MR tools interaction and training as surveyed by Borsci et al. [7].

The following sections will discuss each main evaluation criterion and elaborate on the studies that focus on the mentioned aspects and why they are needed in the evaluation approach.

2.4.1 Technical Aspects and Tool Features

A criterion suggested in academic studies is employed to analyze the technical aspects of training tools, such as system operation, usability, user experience, and satisfaction [9, 65]. These aspects are evaluated through questionnaires given before (for instance, prior to training interaction with the technology) and after the training. The functionality of the tool at various levels (like a low usability level) could significantly influence the content transfer during training by impacting the user's experience with the tool. At the same time, attributes like varying degrees of gamification in training can either enhance or diminish trainees' motivation, presence, and engagement. Therefore, conducting an analysis of these criteria before and after training enables researchers to effectively oversee and understand the impact of these tools on the performance of trainees.

Furthermore, this analysis not only provides insights into the effectiveness of the training tools but also helps in identifying areas for improvement. For instance, if the pre-training interaction reveals a low level of tool usability, measures can be taken to enhance this aspect before the actual training begins. This could involve making the interface more user-friendly or providing additional guidance to the trainees on how to use the tool effectively.

2.4.2 Levels of Acceptance and Participant Attitude

Miller and Kalafatis [43] propose an experiment that guides the user through setting up a piece of equipment, by projecting the digital twin of the equipment and instructions on how to make the assembly steps. They evaluate their system based on efficiency (quickness in task completion), precision (mistakes made during assembly), and complexity (difficulty in following steps). Questionnaires are used to determine the user satisfaction level with the system output with questions related to the level of comprehension (whether instructions are enough and feel complete), and preference (preferred method and which is more intuitive). They discussed the needed improvements after the study and many participants noted that they wanted a means to ask for clarity which would have been less confusing. This is exactly what we are trying to achieve using a dialogue agent. Another limitation they entail is that they mostly used animated holograms, and had limited virtual interactions. We aim to improve this by giving the user the free possibility to interact with virtual models to increase their interactive satisfaction.

2.4.3 Psychological Reactions

Due to the increased complexity in manual workplaces, demand has risen for software solutions providing high practical usability and a low cognitive load on the worker [82]. XR applications can be designed to train specific aspects of cognition such as making problem-solving more efficient [14]. In most studies, these cognitive tasks are in terms of the time taken to reach an objective, finish a specific task, or quantify the cognitive load to project the increase in engagement during the task [7].

The reason why these metrics are important is that with an increase in cognitive performance, higher cognitive levels would mean an easier adaptation to the tasks at hand, making the worker more flexible in adjusting to new technology and completing the training and re-training required in an evolving market [15]. However, in a comprehensive overview of AR assembly training evaluations, Werrlich et al. [82] highlight that even if many literature studies proved faster completion compared to traditional methods, counting the time is an insufficient variable for assembly training, and adding measurements for the quality and training transfer such as immediate recall should be favored when evaluating training systems.

In a study by Werrlich et al. [79], the researchers demonstrate that AR systems outperformed conventional training methods such as video instruction and printed textual instructions in terms of both immediate and long-term recall post-training. Furthermore, it was observed that AR-based training imposed a lower cognitive load compared to traditional training methodologies. The projection-based AR limitations in the field of assembly tasks are also discussed, highlighting the important aspect of complex environments like an engine assembly line and the research gap on how to tackle such complex issues that require many parts.

2.4.4 Level of Immersion

In VR, participants show good progress in immediate assessment after using the headset applications and in further assessments after months, which proves stability in long-term recall. However, compared to AR-based assembly training, in most subjective evaluations VR showed a significantly higher perceived task load and a lower usability

rating when simulating the real workspace [13], while it may also induce cybersickness as a negative effect [56].

Given these drawbacks, while VR software is useful and proves effective in some high-risk training scenarios where participants showed good progress in the immediate and long-term recall [54], the main downside is the inability to see the real world, losing the most important factor in assembling physical objects.

Another limitation to the current evaluation methods that needs to be taken into account is that researchers tend to compare the training systems mostly against paper-based or video-based solutions, and not as much against face-to-face training which is the current training solution in industries [82]. We aim to fill this gap by basing our hypotheses on whether there is a statistically significant difference between face-to-face training and AI-assisted MR training.

Overall, hands-free MR applications with an interactive interface incorporating dialogue agents could be the best solution to increase the cognitive performance of workers and automate the training process in order to make it more efficient and cost-effective. We aim to understand the effectiveness of an MR virtual assistant that can process information the way people do and act as a replacement for a human trainer, which could provide step-by-step guidance for performing tasks, specifying both the nature of the tasks and their spatial context. Evaluating our system will follow best practices from related literature as seen in Table 3, involving a combination of questionnaires for measuring the usability of the system (System Usability Scale [1]), the perceived workload (NASA-TLX), and the acceptance of technology (TAM).

3 Research Question

The research question is stated as follows: How can we develop and evaluate an application for AI-assisted assembly training?

Following this, multiple sub-questions can be derived to tackle the research in a step-by-step manner:

- How can we design an architecture that will take into account the advantages of dialogue agents and MR for assembly?

For this subquestion, the focus is to create an easy-to-use and intuitive training application that works similarly to other software tools that are used in the industry, following the design recommendations and guidelines for creating successful applications with optimal information visualization proposed by Werrlich et al. [80].

- How do we integrate dialogue agents with interactive MR systems such that they are compatible and still highly performant?

This subquestion along with the research material to be used will be expanded upon in the Methodology section.

- How can we evaluate the efficacy of AI-based MR training?

To answer this subquestion, I will use the previous two subquestions to come up with an optimal solution that will be evaluated with a user study on usability testing.

4 Methodology

Based on the findings from related work, we will first identify a specific use case for the system. This involves understanding the problem space, identifying the users, and outlining the tasks the system will perform. Once the use case is defined, we will establish the requirements for the software we are developing. These requirements will be both functional (what the system should do) and non-functional (how the system should perform). The requirements will guide the design of the system’s architecture. The architecture will outline the system’s components, their relationships, and their interactions, providing a high-level view of how the system will be structured and how it will operate.

After determining the requirements, we will conduct a thorough feasibility and risk analysis. The feasibility analysis will evaluate whether it’s possible to develop the system within the given constraints, such as time and technology. It will ensure that the system can incorporate all the proposed functionality and that it can be completed within the required timeframe. The risk analysis will identify potential issues that could hinder the system’s development or operation, such as technical challenges and resource constraints. We will assess each risk’s likelihood and impact, and develop strategies to mitigate them. After this, we will delve into the implementation of the application based on all findings above, which will elaborate the data transfer handling and the integration of the dialogue agent into our system.

Finally, we will evaluate the application through a user study and an experiment. The user study will involve selected users interacting with the system, while we observe and gather data on their experiences. This will provide insights into the system’s usability, effectiveness, and user satisfaction. The experiment will test the system under controlled conditions to measure its performance and reliability. The findings from the user study and experiment will determine the benefits and limitations of our software.

5 Design and Architecture

This section provides a review of the proposed system, beginning with an in-depth use case analysis to comprehend its real-world application contexts. Following this, a rigorous assessment of requirements is carried out to gain a solid understanding of the essential features. Ultimately, the section explores the system architecture, detailing the comprehensive design and framework that will facilitate the project’s successful execution.

5.1 Use Case Analysis

Use case analysis, the foundation upon which the system will be built, is a technique used to identify the requirements of a system and the information used to define the needed processes [28]. The application is meant to be used in the assembly lines of manufacturing companies in which workers assemble components or parts of a larger piece using their hands and a headset that will project digital models that are intractable. Moreover, the users will be guided and supported by a dialogue agent in the form of a virtual assistant that will be able to give instructions when prompted by verbal commands [84]. The interaction with the virtual environment is facilitated solely through hand gestures (poke, grab, pinch), eliminating the need for additional controllers [82]. The intended users are novice trainees with some basic knowledge of the given scenario. This process helps to

design our system from the user’s perspective in order for it to be complete and reach our final goal of satisfying the user [14, 26, 60, 86]. These user needs are organized based on the timeline of user interaction and are structured as follows:

- The user wants to build sets of components without the use of a manual.
- The user wants to visualize the pieces physically and digitally either prior to, during, or after the complete assembly [43].
- The user wants to interact with the 3D Models using their hands, without having to use controllers [79].
- The user wants to receive information about the necessary steps to be taken by asking questions using his voice [84].
- The user wants to hear and see in his field of view the important textual information received [82].
- The user wants to see the correct attachment points emphasized [43].
- The user wants to know when they have completed a step correctly [14].
- The user wants to keep track of his progress by seeing an overview of the steps they have completed and those remaining [81].

5.2 Requirements

This section refers to the analysis of the requirements in accordance with our previous research on what already is in the field of options. The focus is to create an easy-to-use and intuitive training application that works in a similar fashion to other software tools that are used in the industry [14, 43, 81], following the design recommendations and guidelines for creating successful applications with optimal information visualization proposed by Werrlich et al. [80] and the best practices for HMD development [82]:

- **Visual Aids:** Direct visual aids are permanently presented information such as 3D models superimposed on the related real environment. Indirect visual aids are additional information, only presented or available for the user when needed (e.g. text annotations, documentation). This concept allows us to adapt information during the learning process. Clear and detailed instructions at the beginning are necessary for the trainee to understand and perform the tasks.
- **Mental Model Building:** The mental model of an assembly task describes the internal representation of an entire task. Context information such as progress bars can help create a mental model.
- **Passive Learning:** There should be a part in the training where the trainee is not active and only receives information about the task. This concept can help to gain a global picture of the entire task.
- **User Interaction:** The MR experience involves physical manipulation of object components, where the user can ask for instructions to assemble an object. The user interaction should be hands-on without controllers [79], and verbal engagement with the virtual agent should also be possible.

- The final application must show conclusive benefits.
- Performance must be similar to or better than traditional tools.
- User perception and satisfaction must be considered.
- Visual immersion must be a key instruction component.
- The objectives should be well defined, they should not impede a user's self-efficacy or self-confidence, and they must accurately measure the impact on human performance in terms of efficiency, understanding, and accuracy. The user feedback should be immediate and self-explanatory.

Based on the aforementioned recommendations in the literature, the list of functional and non-functional requirements is structured as follows:

5.2.1 User Requirements

- The system must have a list of 3D Model pieces that the user can interact with.
- The trainee must be able to use their voice commands or hand gestures to manipulate the 3D Models and interact with the environment.
- The user must interact with their hands and not with controllers [79].
- The system must have an intuitive interface that will not feel too cluttered for a user.
- The system's interface must provide the user with instructions and responses by the dialogue agent.
- There should be on-demand help during the interaction, with support in either textual, audible, or visual form.
- The dialogue agent should be used to generate information or responses based on the detected states.
- The application could benefit from having a text-to-speech model to provide the user with verbal instructions.
- The final system will not have its own authentication system to eliminate the need to take into account the security of personal information and to prevent data breaches.
- The system will not support multiplayer, as each participant won't have to collaborate with others.

5.2.2 Technical Requirements

- The client-side system shall be implemented in C# using the Unity game engine.
- The server-side system shall be implemented in Python.
- The system shall use cloud storage for storing the dialogue agent's models.

- The system shall store smaller models locally, such as text-to-speech and speech-to-text.
- The system shall use Github for CI/CD and version management.
- The headset must have a good-quality camera in order for the user to see the physical pieces that need to be assembled.
- The headset should be untethered to offer freedom of movement to the users.
- The application should be compliant to existing standards so that it can be easily ported to a new device.

5.3 System Architecture

Creating the system architecture for an MR application can be challenging. However, previous studies have illustrated the essential layers required for MR applications, as depicted in Figure 1. They include the user interface (UI), application logic, and middleware layers. The UI layer is responsible for rendering the virtual objects and managing user interactions. The application logic layer handles the application’s core functionality and controls the flow of the application. The middleware layer functions as a hidden translation layer, enabling communication between the interface and the server. Understanding these layers and their interactions is crucial for the successful design and implementation of an MR application, by guiding the architectural decisions and helping overcome the challenges associated with MR development.

For our scenario, the Unity Engine will shape the architecture of our code base, which incorporates the UI and Application layers similar to Figure 1. In our case, the middleware layer will be the communication between the dialogue agent and the headset. To model the findings for our specific use case, we propose a less generic framework that integrates two major components, the cloud-based dialogue agent and the MR environment within the Unity game engine (Figure 2).

The workflow is structured as follows: the user will ask a question and send an image of their camera feed which will be processed in Unity and sent to the cloud-based dialogue agent. As a response, the AI agent will send the answer to the question and a picture of the manual if needed, which will be displayed on the UI in the headset.

Given the proposed workflow and the requirements described in the previous section, we define the following structure for our system, as seen in Figure 3. The architecture has been designed in a way that allows easy replacement of modules, without disrupting the overall system functionality. This modular approach ensures that each component can operate independently, while still contributing to the collective system goal. Each module and the functionalities they provide will further be described, as well as how they interpolate.

The client-side architecture is shown on the left side of Figure 3, and contains the OpenXR compatibility framework that provides cross-platform support for any XR runtime system such as the Magic Leap 2, Microsoft Hololens, Oculus Quest, or the HTC Vive. The user will then be able to see the Application Interface on any of the OpenXR-compatible headsets (Figure 4).

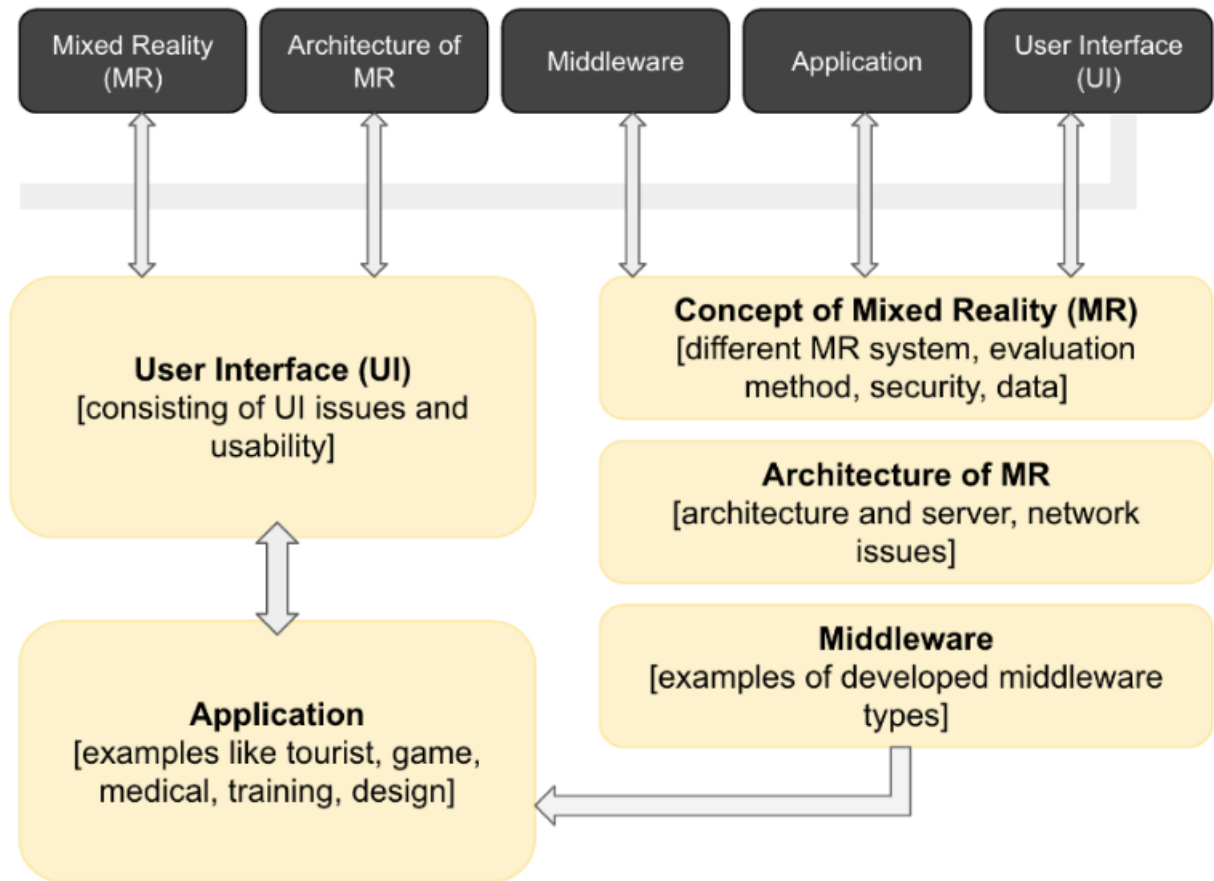


Figure 1. A MR framework split into layers as proposed by Rokhsaritalemi et al. [57].

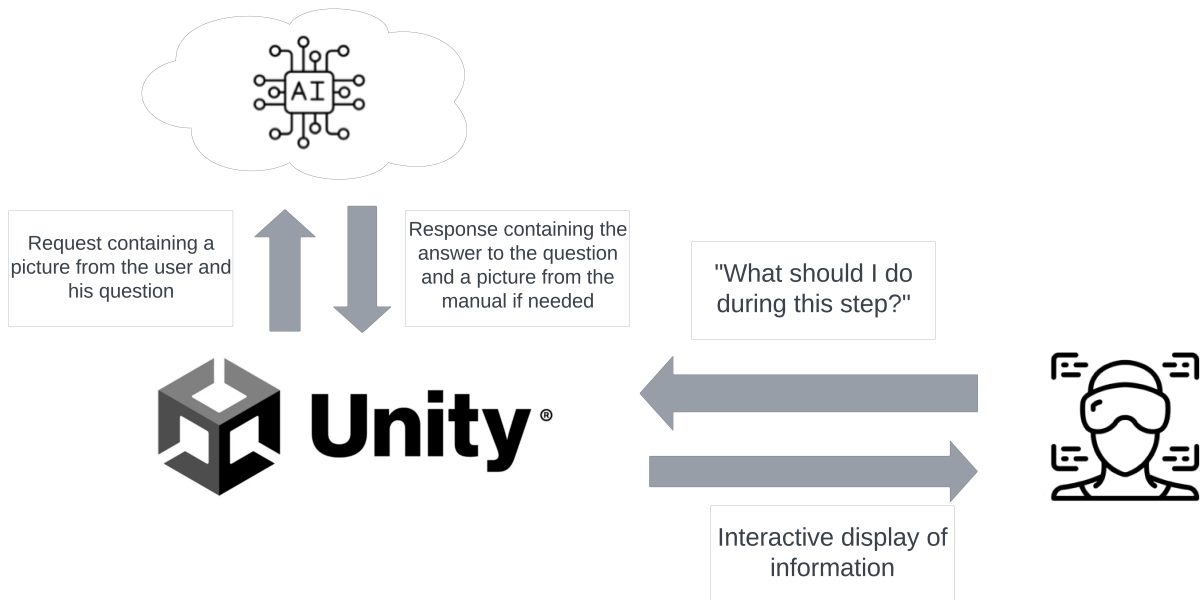


Figure 2. High-level overview for the framework of the system showcasing the data flow between the user and the dialogue agent.

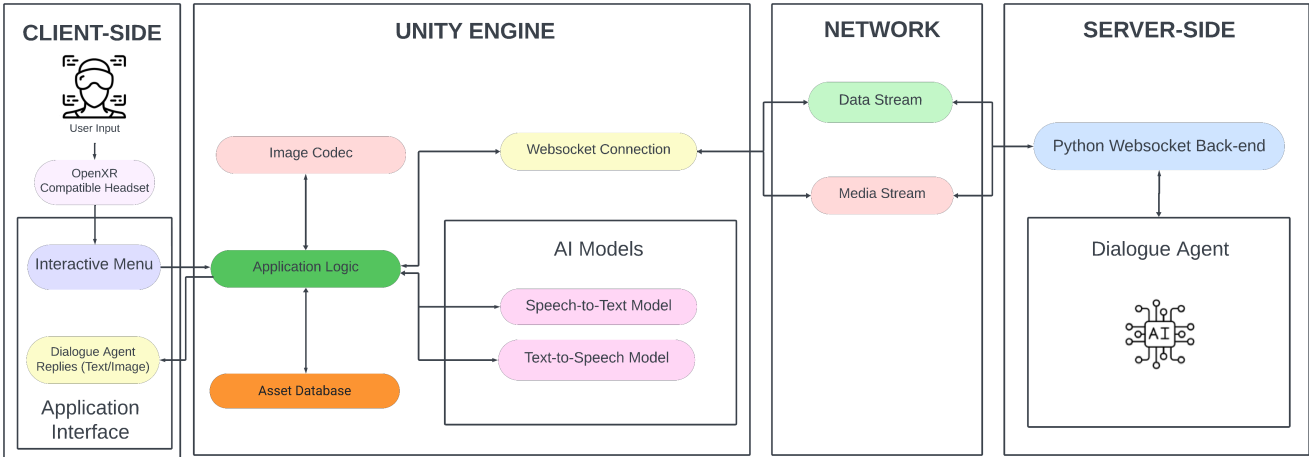


Figure 3. Overview of the system’s components and bi-directional data flow from the client to the server.

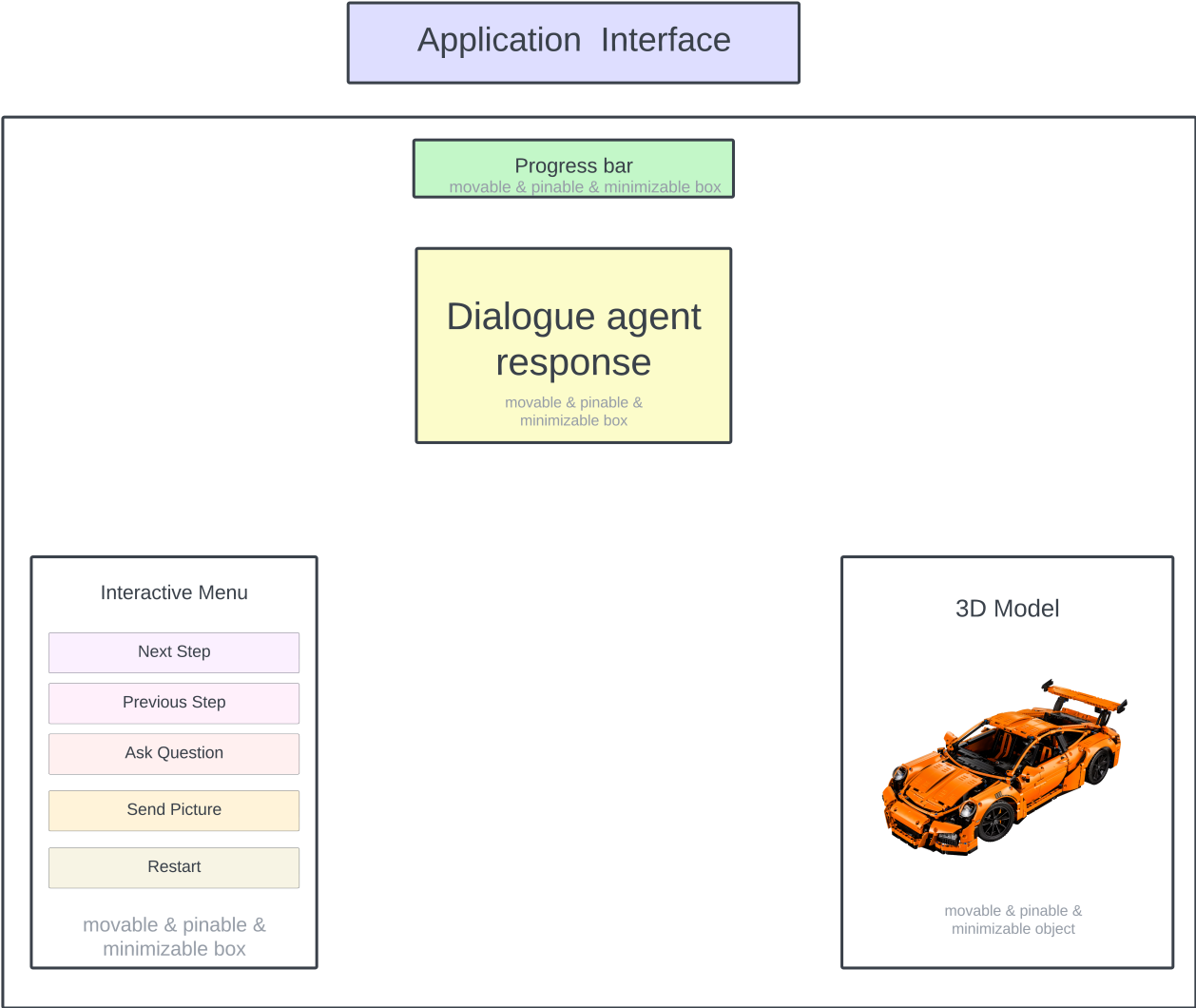


Figure 4. Application interface design containing the interactable dialogue boxes, menus, and models.

The application will be developed and tested on an MR headset, the interface above as well as the overall application design were outlined following best practices for MR development [29, 40, 52]:

- **Start with a home screen.** Starting with a home screen that is familiar and consistent allows users to orient themselves initially. This also helps a user recognize the app’s ”home base” when later using the buttons to navigate [29]. To achieve that, before the main UI, the users will see a coaching UI and video player for onboarding them into the application, providing instructions on interacting with the dialogue agent and the interactive menu.
- **Create simple ways forward and back.** The user needs to have a clear path to follow and to navigate forward and backward through the app [29], in our case by pressing the Next Step and Previous Step buttons.
- **“Billboarding”: orienting objects for readability and usability.** In MR users can view objects from various angles. With billboarding, UI elements will rotate and always face the user, regardless of orientation. This is recommended to increase the readability and usability of text and objects that contain important information [52]. Our UI components will include this functionality.
- **Make actions clear.** Users need to know what, when, and how to choose an appropriate action [29]. If the user chooses to ask a question, he can do it by pressing the Ask Question button. If the dialogue agent responds and asks the user to take a picture in order to get better information, the user can do so by pressing the Send Picture button. Finally, if the user chooses to start the process over, he can choose to press the Restart button.
- **Give adequate feedback.** Since users are required to give input to interact with the application, the right feedback needs to be given when the state changes [29]. We will provide the user with visual feedback in the form of text and images received from the dialogue agent, as well as audio feedback that dictates the sentences received as text.
- **Size and distance for proper depth perception.** If objects are of small size, users can get lost, therefore a best practice is to provide clear and visual-auditory cues when displaying small-sized objects to guide users [52]. The user will have the ability to scale the 3D Model in order to increase its size as much as it is needed. In case the 3D Model is lost in the environment, the user will have the ability to toggle an arrow that points towards the 3D Model.
- **Support direct and indirect input interactions alike.** Direct touch creates an immersive and intuitive experience, designing support for ray-casting from hands lets users interact with objects from a distance [52]. Our application will support this feature such that the menus can be interacted with from any position without the need of controllers.
- **Avoid head-locked content.** Displaying information around the user’s field of view, similar to a heads-up display (HUD), is a popular method for UI design, however in MR the information displayed around the user’s eyes can tire them quickly and reduce usability [52]. The user interface will be fully customizable by the user, with every box and object being movable, pinnable, and minimizable.

- **Design content with sufficient visual and audio cues.** In MR, where virtual objects are integrated into the physical world, it's crucial to design content with ample cues and feedback, given the lack of control over the user's viewpoint. This ensures users can explore and interact with virtual elements without losing their way. Since physical and tactile feedback is absent, visual cues indicating states like "hover" and "pressed" are essential for maintaining quality and usability. When a user focuses on or hovers over interactive objects, visual feedback needs to be provided. Additionally, audio/visual feedback must be incorporated, such as compressing movement or highlighting upon pressing, accompanied by audio cues. [40, 52]. When hovered, the buttons of the application will provide an audio sound to signal the user that they are selected, and when the user presses the button, the color will change to fill it based on the depth it has been pressed. Once it is fully pressed, audio feedback will be given to notify the user of the change.
- **Design experiences to be spatially responsive.** A user's flow through space needs to be taken into account, as they need an adequate and clear environment to interact with the interface [40, 52]. The application will consider a viable assembly environment, with minimum space of a desk/table as the surface needed to assemble the model, while the UI will be projected similarly to a digital workplace [72].

Following established best practices, the default interface will not be head-locked; instead, it will appear in front of the user's initial position. For instance, if the trainee is seated at a desk, a prompt from the dialogue agent will be displayed in the upper center of their field-of-view (FOV), the real model will be in the bottom center, and the menu along with the virtual 3D model will be positioned on the sides. The interface is designed to allow the user to see their current progress, the dialogue agent's responses when needed, the current step of the 3D model of the Lego set (a digital representation of the physical Lego they have assembled up to that step) and an interactive menu with all available options. Each UI element can be dragged, moved, and pinned wherever the user prefers. Each element will also have a button to minimize the specific box. This flexibility allows the trainee to customize the interface layout and remove any elements that are unnecessary during specific assembly steps.

In the Application Interface, there are two modules, the Interactive Menu, and the Dialogue Agent Replies. The Interactive Menu facilitates user interaction with buttons, transmitting these interactions to the Application Logic. Meanwhile, the Dialogue Agent Replies module receives text/image-based responses from the server via the Application Logic.

The Interactive Menu module includes various UI components such as buttons, progress indicators, and slates. These elements are part of the Mixed Reality Template, a starting point for MR development in Unity [70]. This template is designed to speed up the creation of MR and AR applications, offering a cross-platform input system and foundational elements for spatial interactions and UI. For instance, the menu in the bottom left of Figure 4 is a control element containing an array of buttons and other UI components. The template also features the General Grab Transformer script, enabling the 3D model in the bottom right to be moved, scaled, and rotated using one or two hands.

The Application Logic module within the Unity Engine serves as the central controller for other modules based on the application's state. The main functionalities are defined as follows:

- Receives user interactions from the Interactive Menu module and returns the adequate data depending on the application state in the Dialogue Agent Replies module (next/previous 3D Lego piece requested and retrieved from the Asset Database, as well as text, images, or audio responses from the Dialogue Agent).
- Creates and manages the bidirectional connection between the client and the server using the WebSocket protocol, ensuring seamless and efficient communication for exchanging real-time data and messages. This involves handling the creation of connections, managing data transmission in both directions, maintaining connection stability, and handling any errors or disruptions that may occur during the communication process.
- Establishes bi-directional communication with the Image Codec module to transmit and receive images, facilitating their encoding/decoding and subsequent transmission/reception to/from the user/Dialogue Agent via WebSocket connections. For instance, when a user captures an image through the headset camera, it is called within the Application Logic, where it undergoes encoding using the Image Codec module. Afterward, the encoded image is transmitted over the network through the Media Stream module. Similarly, images received from the Dialogue Agent are directed toward the user, following the opposite direction of the communication flow.
- Facilitates bidirectional communication with the AI Models to handle the conversion of speech to text or text to speech. If the user chooses to press the button that sends a question to the server, the user's speech will be transmitted by the Application Logic to the Speech-to-Text Model (within the AI Models block). The Application Logic will then receive the text converted from speech in a string format. The Application Logic proceeds to transmit this text to the server through the Data Stream module using the WebSocket connection. Upon receiving a response string from the server via the WebSocket connection, the Application Logic transfers it to the AI Models. The Text-to-Speech Model then transforms the text into phonemes, which are sent back such that the resulting audio output is played in the user's headset.

The Asset Database module serves as an essential component within the Unity Engine framework. Its primary responsibility lies in efficiently managing and providing access to a wide array of resources, including 3D Lego pieces, as requested by the Application Logic. This module ensures seamless integration and retrieval of assets within the Unity environment, enabling the Application Logic to dynamically acquire and utilize the necessary 3D elements for various functionalities and interactions within the application (such as the UI components).

The AI Models play a pivotal role in the application's functionality, specifically in the interpretation and generation of speech. The Unity Engine offers a comprehensive framework known as Unity Sentis, which enables the seamless integration and execution of AI Models directly within applications [69]. Sentis harnesses the computational capabilities of end-user devices rather than relying on cloud infrastructure. This approach eliminates the need for complex cloud setups, minimizes network latency, and eliminates recurring costs associated with cloud-based inference.

The AI is responsible for decoding the user’s spoken words, a task accomplished through the utilization of a Speech-to-Text model, leveraging the capabilities of Generative AI. This model processes the audio input, accurately transcribing it into textual format, which is then relayed to the server for further processing. When the dialogue agent formulates a response, it generates a textual output. The AI Models then undertake the conversion of this text into phonemes, a process handled by a Text-to-Speech model. This model incorporates an embedded dictionary to translate the textual representation into phonetic units, which are synthesized into audible speech.

The Image Codec module in Unity Engine is utilized by the Application Logic to prepare the images for transmitting the media to the Dialogue Agent or the user’s headset. For instance, when a user captures an image using the integrated camera, the Image Codec encodes the image and forwards it as a request to the dialogue agent. Upon receiving the agent’s image response, the encoded data is decoded before presenting it to the user. It’s worth noting that the display of most headsets and their camera have differing FOVs. This discrepancy means that the display can render content that is vertically larger than what the RGB camera can capture. Consequently, it is imperative that the capturable area is of the right dimensions before a user captures an image. This ensures that the physical model remains intact and prevents it from being cropped when transmitted to the dialogue agent.

The WebSocket Connection Module creates a bidirectional communication between the server and the client. The module initiates the WebSocket handshake process, allowing the server and client to establish a persistent, full-duplex communication channel. During the handshake, the server and client agree on the WebSocket protocol version to be used for communication. Once the connection is established, both the server and client can send and receive data asynchronously. This enables real-time data exchange between the two parties.

Over the Network, segmenting data and media streams into separate channels, such as a dedicated Data Stream for textual information and a distinct Media Stream for handling images, offers significant advantages for WebSocket-based applications. This approach enhances organizational clarity, streamlining data management and providing easier navigation within the application architecture. By optimizing channels for specific content types, it boosts overall efficiency and performance, ensuring swift transmission of text and seamless rendering of images.

The server-side implementation, situated on the right side of Figure 3, comprises a WebSocket backend responsible for the data exchange with the client. Incoming user requests, transmitted over the network in either text format (Data Stream) or as encoded images (Media Stream), are directed to the Dialogue Agent. The Dialogue Agent processes these requests and formulates appropriate responses, which are then relayed back to the user via the WebSocket connection. This setup enables seamless communication between the client and server, facilitating real-time interaction and dialogue.

The system architecture provides a high-level overview of the entire system, including the main components and their interactions. On the other hand, a control flow diagram displays a more granular view of the system, showing the order of operations and how the information flows between different parts of the system (Figure 5).



Figure 5. Control Flow Diagram describing the functionality of the system.

When the application starts, it establishes a WebSocket connection between the headset and the server and after it is tethered, the server sends a welcome message to the user, with instructions for interacting with the dialogue agent and the menu, as seen in the best practices of MR development. If the connection fails, an error message will be displayed to the user to check the connection to the internet. After the welcome text box is dismissed by the user, they will digitally see in the 3D model box of the Application interface, the first piece that he needs to find in the Lego set, to start the model assembly.

Once they find the piece and press the Next Step button, the Application Logic module will fetch from the Asset Database the next piece that they will need to assemble, which will be highlighted and positioned in the Lego model (displayed on the 3D model box in the UI), connected to the previous digital representation of the piece.

If the user chooses to press the Previous Step button, and there is a previous step, it will display the prior digital piece highlighted and positioned in the Lego model. Otherwise, the first step of the building process will be displayed. The user always has an option available on the UI to restart the operation within the application. If the Restart button is pressed, the first piece of the 3D model will be displayed on the interface.

During the assembly procedure, the user can request assistance from the Dialogue Agent. If he presses the Send Picture button to capture the physical environment using the attached headset camera, the picture will be sent to the server, while the server will send back instructions to the user with textual, audio, and visual assistance, based on the current step in the building process that was received from the user. The Dialogue Agent will explain the current step, and it will be displayed to the user as text (which is also translated to speech as audio feedback), along with a picture with arrow signs indicating where the current piece needs to be located in the built model. If the user presses the Ask Question button, the microphone of the headset will be activated for them to receive their question. The speech will be converted to text and sent to the server. As in the previous case of the Send Picture button, the instructions from the Dialogue Agent will be displayed on the UI.

If there are no more next steps, the build will be completed and the user can see the full version of the digital 3D Model on the interface.

6 Development

This section presents the proposed development of the application. A feasibility study is undertaken to evaluate the project’s feasibility from technical, operational, and time perspectives. The analysis further includes risk assessment, pinpointing potential obstacles, and formulating strategies to counter them. Finally, the last section is focused on detailing the implementation.

6.1 Feasibility Study

A feasibility study is used to decide whether, given a list of requirements and a certain amount of time, the project can be completed, based on three main aspects. This study was performed after finalizing the requirements for the final application. The following sections will elaborate on each of the three parts that the study is built upon, in order of appearance: technical, operational, and time feasibility.

6.1.1 Technical Feasibility

To determine the technical feasibility of the project, the technical decisions made, regarding what will be used to approach the project, are discussed together with complications that could follow due to these decisions. The aim is to build the next generation of personal assistants and to conduct research towards improving human-to-machine interaction. The application is meant to engage and inform those new to assembly tasks, as well as those who already know how to perform them well. The trainees would be able to interact with 3D models using their hands as well as voice commands.

The needed technology is based on the discoveries made from the related work, which means that for hardware we require a headset that can support MR and user interaction without the need of controllers (hands-free support). We aim to use either a standalone Oculus Quest 3 headset or the same headset tethered to a computer and the reasons will be explained in the next section. We will also require a server that will host the AI dialogue agent. No additional equipment is envisioned to be necessary for our experiment as all interactions within the training environment occur with virtual holographic content and audio information.

Regarding software, we use the C# based Unity engine to shape the architecture of our code base. Unity Engine was used as it allows us to easily create graphical applications that can load models and show data in a clear and well-organized fashion.

6.1.2 Operational Feasibility

Operational feasibility is a measure of how well a solution meets the identified system requirements to solve the problems and take advantage of the opportunities envisioned for the system [6]. To judge the operational feasibility of the project we have to take a look at what is required to solve the problem at hand, our application and compare it to the current trend of development. There are 2 types of XR technologies on the newly revised continuum: AR (MR being a subset) and VR. Looking at the accessibility of these technologies right now, AR is the most accessible as it works with our smartphones, which means it is widely spread and easily attainable.

Most AR research is currently conducted using HoloLens 2 headsets due to their widespread use in businesses for creating step-by-step visual work instructions. However, with the discontinuation and end of support for HoloLens 2 [41], exploring alternatives like the Oculus Quest has become necessary, despite its primary use in VR. A study comparing the Quest 2 to the HoloLens 2 concluded that the Quest is too constrained to serve as a viable alternative, mainly due to its lack of features such as object recognition, LiDAR, a depth sensor, high-resolution cameras, and access to the camera feed [10] [5].

However, recent advancements in MR have led Meta to release the Quest 3, which includes dual 4MP cameras and a depth sensor (still only half the quality of the HoloLens' 8MP cameras) [53]. Along with this release, the Passthrough API has been updated, enabling developers to create simple MR applications using Quest headsets. These headsets also have a microphone and speaker, which are beneficial for providing a dynamic language experience with a virtual assistant during training sessions. We will be using the Oculus Quest 3 since Utrecht University has one readily available for experiments.

One significant challenge in developing our application on a Quest device is Meta's restriction on capturing or streaming images or videos of the physical environment through the Unity Editor due to privacy reasons. The Passthrough feature is rendered by a dedicated service into a separate layer, where the cameras generate a sparse 3D point cloud. This layer reprojects the camera views to match the depth of this point cloud, and then color is added using the central color camera. To stream the image and perform object detection, a standalone version of the application needs to be built inside the headset, or a screen copy of the Unity Editor must be projected on a computer and then sent to the dialogue agent. As mentioned, a computer will likely be used to host the server-side MR application, which will communicate directly with the headset.

To demonstrate the capabilities of the application in assembly tasks, the best approach would be to experiment with real workers on an assembly line [81]. However, since it is beyond the capabilities of the master thesis we will be focusing on a similar objective that we can control which entails completing a training scenario offering Lego set-building assembly training through the use of mixed-reality HMDs, as seen in previous studies [13]. Participants will engage in assembling Lego pieces using their hands and with the help of digital 3D models that can be interacted with [43].

6.1.3 Time Feasibility

An important factor to consider in the feasibility study is the time required to complete the project. With about 40 hours of work per week, this is deemed reasonable given the 5-month deadline set by the master's thesis. Initially, estimating the time needed to implement various features was challenging due to limited experience in developing a major experimental application and working with MR software. Nonetheless, considering the time frame and the experience gained from previous projects, the requirements are feasible and achievable.

6.1.4 Summary of Feasibility

In summary, since all three aspects of feasibility discussed above are achievable, the project as a whole is feasible. Given the schedule, time constraints, and available resources, the project and all implementation steps are technically, operationally, and temporally feasible.

6.2 Risk Analysis

Risk analysis is the process of identifying and analyzing potential issues that could negatively impact key project goals. Research needs to be conducted about these issues, identifying them, and coming up with some potential solutions. In order for the project to be developed efficiently, with no incidents, it is important to know about what issues may arise and how we might solve them.

We would like the final application to work on any type of headset. Porting an app from one device to another is not always a simple process, there may be some challenges and limitations that need to be considered. Firstly, the lack of experience with programming software used primarily on headsets poses a risk, because of the unfamiliarity with technologies and frameworks. Secondly, unlike computer applications, the user’s ability to roam around with all degrees of freedom needs to be taken into account.

Another potential risk is the limitations of the hardware used to build the application. Having a low-resolution display would make interacting with real-world objects much harder [10]. In our current context, users might not be able to see small Lego pieces, so in order to mitigate this, if it is a visible issue after the prototype implementation, larger pieces should be preferred instead.

Working with LLM-powered dialogue agents presents storage challenges, often requiring cloud hosting. This leads to costs associated with network traffic for every API call during debugging and experimentation. Additionally, latency is a concern with AI models, though smaller models that don’t need cloud hosting can help mitigate this issue. Recently, Unity released an experimental feature called Unity Sentis, which allows AI models to run directly on user devices through the Unity Runtime [69]. This approach leverages the computing power of end-user devices, eliminating the need for complex cloud infrastructure, network latency, and recurring inference costs for tasks like speech recognition or object detection.

Finally, introducing such a system as a method for training should take into account an array of factors, such as privacy concerns related to streaming the room surroundings through the headset camera and the sending of user input over the network.

6.3 Implementation

The implementation phase involves translating the design and requirements into a functional MR application. This section will detail the steps taken to bring the project to life, including the development process and integration of various components, as well as the final application walkthrough.

The primary challenge is organizing and managing the user’s input along with the textual and visual instructions provided by the dialogue agent. Unity acts as the bridge between the physical and virtual worlds, facilitating seamless data transfer between them. Given that the AI agent is multimodal, the process involves converting the user’s speech input into text and capturing their current view as a 2D image, which the user manually sends over the network. Subsequently, the user receives a response consisting of a text-based answer and a 2D image with visual instructions. Data is transmitted to Unity via a Python-Unity socket, ensuring two-way communication between the Python script and the Unity application. In this setup, Unity hosts the client, while the Python scripts operate the server. The user can then view the dialogue agent’s response and decide the next steps to complete the assembly task.

6.3.1 Technical Details

Our tool is built upon the MR Template [70], which can be selected when starting a project in Unity. This template includes the necessary packages for setting up an MR application, such as gesture interactions and UI elements. The template already has preinstalled the packages needed to support development on OpenXR platforms, as well as gesture interactions and standardized UI elements. In addition, the Affordance system provides feedback for the user with visual and auditory cues (e.g., when pressing buttons). This requires the use of the XR Interactable Affordance State Provider with a specified interactable source.

The 3D Lego models of a Lego Dump Truck and a Lego Truck Cabin used for the experiment were created in Blender using online Lego brick models [68], scaled to match the size of the physical Lego models. There are a total of 19 pieces per model. The physical Lego pieces were purchased from an online store. The manuals for the models were created using pictures of the 3D models, accompanied by text describing the pieces to be assembled. The models have an XR General Grab Transformer and an XR Grab Interactable component that allows the user to grab, rotate, and scale the object using either a direct or ray interactor.

Initially, the application client will connect to the server hosting the dialogue agent. This involves configuring the IP and ports on both the client and server sides to send and receive data. Once this configuration is complete, both the application and the server need to run simultaneously. The trainer will then select the model to be assembled and the microphone to be used during the assembly. After these steps are completed, the headset can be handed over to the trainee.



Figure 6. The start screen of the application which includes the 3D Model to be assembled as well as the onboarding instruction cards.

The spatial UI enables both near and far interactions with UI elements and includes a coaching UI to onboard users into the MR application, as shown in Figure 6. These onboarding cards guide users through the features, providing instructions on how to interact with the menu and the dialogue agent, as well as how to grab, rotate, and scale the model. The Coaching UI GameObject is managed by the Goal Manager utility class within the MR Interaction Setup. The Goal Manager acts as the main controller of the application, handling the Application Logic as depicted in the System Architecture diagram (Figure 3). It oversees the progression of content within the UI, toggles related GameObjects on and off, and adjusts the Lazy Follow behavior of the UI based on the instructions for each step.

After the onboarding goals are either skipped or completed, the user will see the Interactive Menu, the 3D Model, and a panel containing input instructions (Figure 7). These UI elements adhere to the standards used in the default UI of Meta headsets. The Tutorial Video Panel GameObject within the UI includes a video player that demonstrates basic input mapping using pinching. To the top right of the panel, there is a button that once pressed, showcases the second video explaining poking interaction. Users can move the canvas in space by grabbing either the header or the handle at the bottom of the canvas. To adhere to the best practices discussed in Section 5.3, the billboarding component is enabled on the prefab by default, with the canvas's positional transformation determined by direct/ray interaction. This functionality is also utilized in the Interactive Menu Manipulator and Helper Menu Manipulator GameObjects, allowing users to interact with the model and the dialogue agent.

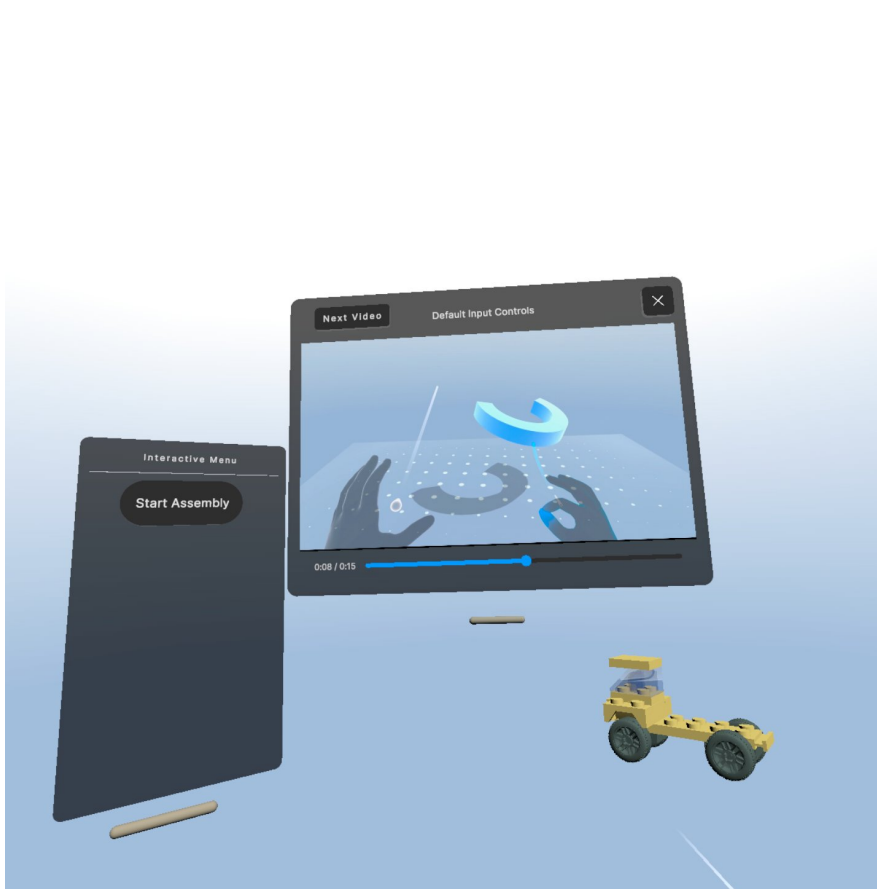


Figure 7. User view after completing the onboarding cards. They will see the interactive menu to their left, the 3D Model to the right, and the input instruction videos towards the middle.

When the user presses the 'Start Assembly' button, the 3D model will then display the first piece to be assembled. The tutorial videos will disappear (unless already closed by the user), and a progress bar along with a textual response from the dialogue agent will appear in their place to display the main UI (Figure 8). The user can navigate through the assembly steps using the 'Next Step' or 'Previous Step' buttons. The NextStep() and PreviousStep() functions within the GoalManager utility class handle the progression of the 3D Model through the steps, update the elements, and perform the color animations.



Figure 8. The main UI containing an interactive menu that displays all the interactable buttons, a progress bar as well as a text prompt from the dialogue agent.

Interaction with the dialogue agent is initiated by pressing the 'Ask Question' button. Upon pressing, the button text changes to 'Stop Recording'. If the user does not manually press the button again to stop the recording, it will automatically terminate after 10 seconds. The AskQuestion() method handles recording the user's question through the microphone and sends it to the AI model for transcribing the given audio to text. These models are stored within the StreamingAssets folder, Unity's default location for runtime data loading. This approach ensures models are loaded only when needed, conserving memory and minimizing performance impact. After converting the speech into text, the resulting string is sent to the Application Logic's GoalManager, which starts a coroutine to transmit the question to the dialogue agent. This involves sending the text over a WebSocket connection as part of the Data Stream.

When the question reaches the server, the dialogue agent processes it and returns a text response. On the client side, the `UdpSocket` class takes the textual answer and dispatches it to the main thread to be transformed into speech. The script loads the necessary text-to-speech model and phoneme dictionary, performs text-to-phoneme conversion, runs the AI model inference, and plays the generated speech audio. Additionally, the text is displayed in the dialogue box on the UI.

If the trainee's question lacks sufficient information for the agent to provide a response, the dialogue agent may request additional visual data. In the scenario present in Figure 9, the dialogue agent cannot establish by itself the current step of the user and requests a picture to be sent back with the real-world Lego set.

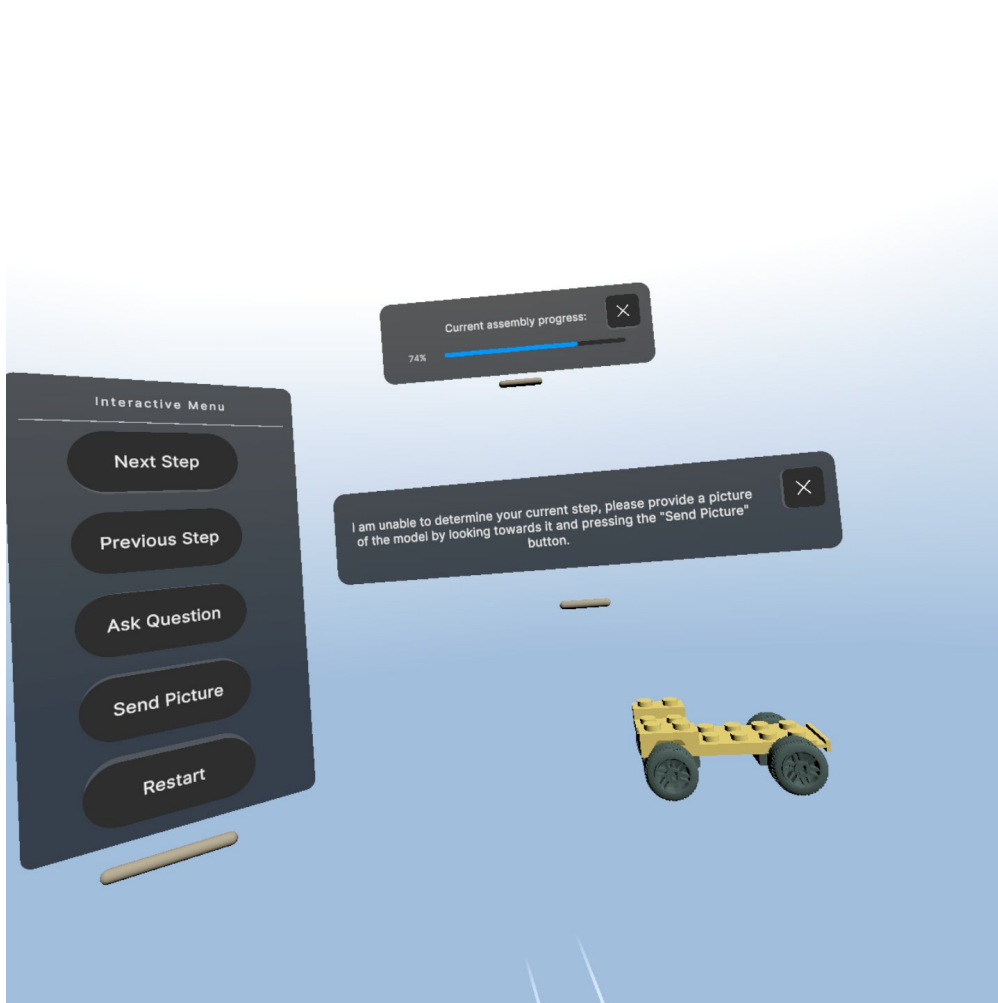


Figure 9. User view after pressing the 'Ask Question' button. In this case, instructions are given within the dialogue agent text box to send a picture of the real-world model.

When the 'Send Picture' button is pressed, the `SendPicture()` method within the `GoalManager` initiates a coroutine to send the picture to the dialogue agent. Once activated, the button text changes to 'Please wait...', indicating to the user that the picture is being transmitted over the network. The coroutine starts by capturing the screen from the trainee's current view. Due to the high resolution of the camera, the default screen capture results in a very large file, which is too big to send in a single packet to the server. To address this, the screenshot is first encoded to JPEG, a format known for its lossy compression that reduces file size by discarding some image details. The compression

level is adjustable, balancing image quality and file size. Currently, the compression is set to medium quality (50 on a scale from 0 to 100), providing an optimal balance to ensure the AI model performs well in object recognition while keeping the file size manageable. Depending on how well the AI detects the current step in the Lego set assembly, the compression level can either be set higher or lower.

When the encoding process is completed, the packet's chunk size is specified, currently set to 8192 bytes. On average, this results in about 20 chunks per screen capture. The image is segmented into these chunks, encoded into base 64, and transmitted over the network. Subsequently, the server verifies if all chunks have been received; if not, it requests the client to resend the image. This is necessary due to potential packet loss during transmission over a WebSocket connection. After receiving all packets, the base 64 encoded image data is extracted and decoded into bytes. These decoded chunks are accumulated into a byte array, allowing for the reconstruction of the image on the server side.

Once reconstructed, the dialogue agent can ascertain the current step of the trainee in the assembly process. Having identified the current step, the dialogue agent prepares a response to send to the client side. This response includes instructions from the manual presented in both text and image formats, depicted in Figure 10.



Figure 10. User view after pressing the 'Send Picture' button. The dialogue agent sends a picture from the manual as well as textual assembly instructions.

The textual response adheres to the structure used with the 'Ask Question' button. Similarly, the image response replicates the chunk reconstruction process implemented on the client side. On the server side, the Python code mirrors the functionality of the C# functions implemented in the Unity client. Images are transmitted over the socket as byte chunks encoded in base 64. Once all packets are received on the client side, the UI is updated to display the dialogue agent response, Moreover, the text is converted into speech to provide auditory feedback to the trainee.

Once the Lego set is assembled, pressing the 'Restart' button will reset the assembly process to its initial state. For data logging purposes, which are used for result analysis, the Restart() method also includes functionality to send the server data regarding the number of UI elements pressed and interactions with the 3D model.



Figure 11. *Additional Helper Menu containing toggles to assist the user in case something goes wrong during the experiment.*

Positioned 90 degrees to the right of the user's main view is an additional helper menu, intended for use if something goes awry during the experiment (Figure 11). If a button is accidentally pressed during the initial onboarding, the user can use the relaunch button on the helper menu to revisit the welcome instructions. Additionally, the user can switch between VR and MR by toggling the passthrough feature, allowing them to view the digital 3D model directly in VR instead of the real world. This feature is also useful for relocating the model if it is accidentally lost during assembly. The progress bar can be toggled on and off to manage the amount of information displayed on the UI. The

tutorial videos for basic input mapping, such as pinching and poking, can be replayed using this menu. Lastly, there is a toggle for a model finder, which activates an arrow pointing toward the 3D model if it becomes misplaced in the assembly process.

To integrate the latest Unity features, including Unity Sentis, which allows for the use of built-in AI models at runtime without the need for API calls, the project is currently utilizing Unity 6 Beta version 6000.0.0b16. In addition to the packages from the Mixed Reality Template, the project employs several external libraries: TextMeshPro, Newtonsoft JSON, Sentis, and Demigiant DOTween. Development was conducted using a Meta Quest 3 Advanced All-in-One VR Headset, which boasts a resolution of 2064x2208 pixels per eye, a 120Hz refresh rate, a maximum FOV of 104 degrees, a Qualcomm Snapdragon XR2 Gen 2 processor, 8 GB of RAM, an IPD range of 58 to 71mm, and 6-DoF Inside/Out tracking. The device also features two 4MP front-facing cameras with full-color pass-through capability.

The enlarged versions of the figures from this chapter can be found in Appendix A.1. The codebase is publicly available via the following link: [GitHub Repository - MR Tool](#). The repository includes a README file that provides a comprehensive overview of the project structure, along with detailed descriptions of the main classes that contain the core functionality, and how to connect the server to the application running on a headset. For any inquiries regarding the setup or use of the application, please feel free to reach out to Luca Becheanu at luca.becheanu@students.uu.nl.

6.3.2 Application Walkthrough

At the beginning, the trainer selects the model that needs to be assembled, as well as the microphone to be used during assembly. After this initial step, the headset can be given to the trainee which will see four instruction cards detailing the experiment setup and how to interact with the environment. Once the cards are either skipped or read, the user will see the real world and his UI. Before starting the assembly, there are two tutorial videos explaining pinching and poking. To the left, the interactive menu can be found, while on the right lies the 3D Model. Further to the right there is another helper menu to be used just in case something does not go as intended. It contains a toggle for the tutorial videos, the onboarding cards, the progress bar, the passthrough (the switch between VR and MR), and a model finder (an arrow that points toward the 3D Model in case it gets lost during assembly).

Once the user presses on the 'Start Assembly' button, the tutorial videos will disappear (if they were not already closed by the user) and a progress bar as well as a textual response of the dialogue agent will appear in their place. The 3D Model will now show the first piece that needs to be assembled. The user can press either the 'Next Step' or 'Previous Step' button to change the assembly step. The dialogue-agent interaction can be done with either the 'Ask Question' or 'Send Picture' button. Finally, the Restart button will reset the assembly to the beginning. Since the application is highly dynamic, the results are best visualized inside a video, rather than in figures. For a visual walk-through, the following link is available: [Multimodal Immersive Systems for Assembly in Mixed Reality](#).

7 User Study

This section focuses on the investigation of variables, hypotheses, experimental setup, tasks, and participants. It starts by identifying and classifying variables that are significant to the research. Hypotheses are then developed to predict the expected results based on these variables. The experimental setup is elaborately explained, providing details about the environment, equipment, and conditions under which the research will take place. The tasks that will be carried out during the experiment are specified, clarifying the particular actions or processes that participants will engage in. The subjects involved in the study are described in terms of selection criteria. Finally, the results will be analyzed, and the evaluation will ascertain whether the hypothesis is supported.

7.1 Variables

Many variables are involved in the experimental process, and some variables have to be controlled to prevent the influence of the results. Those variables are divided into different types such as control variables, independent variables, and dependent variables.

The control variables are not the subjects of the research, but they can affect the experimental results. For example, different devices have different capabilities based on their hardware. A low-quality camera from some HMDs could influence the usability of the application as the pieces would be harder to distinguish. When performing our experiment we will be evaluating our study on the same device. In the case of the independent variable, we will focus on the instructions given by a person with a manual compared to the instructions given by a dialogue agent. Finally, the dependent variables are performance (how fast a person goes from the initial state to the end state), quality (measured by mistakes made during assembly), the number of questions asked, and cognitive load (with questionnaires).

7.2 Hypotheses

Based on the related work highlighting the research gaps, this study aims to evaluate how well AI-based MR can be used to train a user in performing assembly tasks, as compared to the traditional following instructions from physical manuals and face-to-face training. We propose the following hypotheses:

H0: There is no statistically significant difference between following instructions from face-to-face training and AI-assisted MR training.

H1: There is a statistically significant difference between following instructions from face-to-face training and AI-assisted MR training.

7.3 Experiment Setup

To ensure our experiment would not be affected by other factors, the tasks are performed in a private space, within the rooms provided by Utrecht University. The participants are briefed on the tasks they have to accomplish and they are provided with instructions on how the application works. The training will focus on a single user participating in the environment at a time.

The experiment follows a within-subjects design, in which all subjects test both conditions of face-to-face training and AI-assisted MR training. Although the second assembly attempt is inherently biased since participants have completed a similar task before, this setup is chosen such that the participants can provide feedback on both methods and determine which method they would rather use in an assembly scenario. The order in which the methods are presented will be counterbalanced across participants to mitigate potential biases that might arise from the sequence in which the tasks are performed. The participants will be randomly assigned to one of the four possible combinations of training methods:

- Using the MR tool to assemble a Lego Dump Truck, followed by face-to-face training to assemble a Lego Truck Cabin
- Face-to-face training to assemble a Lego Dump Truck, followed by using the MR tool to assemble a Lego Truck Cabin
- Using the MR tool to assemble a Lego Truck Cabin, followed by face-to-face training to assemble a Lego Dump Truck
- Face-to-face training to assemble a Lego Truck Cabin, followed by using the MR tool to assemble a Lego Dump Truck

To address any partiality, we will use Lego sets with different steps and a few different pieces for each assembly task. This variation helps ensure that participants are not simply repeating the same process, providing a more accurate assessment of each training method’s effectiveness. The insights gained from their feedback will highlight what aspects worked well, suggest areas for future improvements, and identify the system’s limitations.

An Oculus Quest 3 is used to immerse the user, with the application running standalone on the headset. The standalone nature of the headset allows users to move freely and interact naturally with the virtual components, making the assembly process more enjoyable and effective. A computer hosts the server where the dialogue agent resides. The user receives support through textual information and pictures indicating where to insert the parts for the current step. This process continues until all parts of the Lego set are assembled.

Unfortunately, at the time the experiment took place, we did not have access to an AI dialogue agent trained on a Lego dataset that could provide the functionality we require, therefore we will be using the Wizard of Oz approach in our experiment [38]. The subjects will interact with the system believing it is autonomous, but instead, the server side of the system is partially operated by a human being (in this case the trainer, as seen in Figure 12).

Modifications have been implemented in the Python script to manage questions directed at the dialogue using the keyboard instead. The trainer can type a response directly into the server’s console when a user asks a question. If the question relates to the user’s current step, the dialogue agent (the trainer) will prompt the trainee to send a picture of the Lego set. When the trainee presses the ‘Send Picture’ button, the updated Unity C# script also sends the current step of the virtual 3D model over the network. This serves as a workaround for not having an AI model capable of performing object recognition.

The trainer then has the option to send the instructions for the current step from the manual or determine if the user is stuck on a previous step. If the user is behind, the trainer can send the relevant text and image for that earlier step instead. This ensures that the user receives the appropriate guidance based on their progress and any difficulties they may encounter.



Figure 12. A trainee conducting the experiment is scaling a virtual 3D model. The laptop at the bottom of the picture runs the Python server hosting the dialogue agent. The trainer responds to the trainee’s questions by typing answers into the server’s console.

7.4 Tasks

Before the experiment, the subject will fill in a consent form with information about the participant’s rights, and a short briefing on the task they have to accomplish, followed by a demographics form (Appendix A.2.1). Afterward, the two Lego set-building training methods will be investigated, one with the MR tool, and the other with face-to-face training. The participant is assigned one of the methods. If the MR tool training is assigned first, the user will wear the headset and complete the pre-training setup, which includes reviewing the onboarding cards and video tutorials, as well as customizing the interface to meet their preferences. The trainee will be given as much time as needed to familiarize themselves with the MR tool, practicing button presses using poking or pinching gestures.

Once the pre-training is complete, the user will press the 'Start Assembly' button, initiating a timer that will stop when the Lego set is fully assembled. During the MR tool training, various metrics such as button presses, the number of times the digital Lego model is grabbed, questions asked, and mistakes made during assembly are logged for further analysis. The trainee will follow the assembly steps using gestures to interact with the virtual 3D model, and they can ask the dialogue agent questions to receive verbal instructions and pictures from the manual if needed. After completing the assembly, the trainee will fill out a Tool Usability Form (Appendix A.2.2). This form is based on the Technology Acceptance Model and the System Usability Scale questionnaires but includes perceived usefulness questions tailored for assembly scenarios. The final questions of the form are taken from the NASA-TLX questionnaire to assess the mental workload experienced during the assembly.

With the first part of the experiment complete, the trainee will assemble the second Lego set with the assistance of a human trainer. Once again, the time taken to complete the assembly, any mistakes made during the process, and the number of questions asked are recorded. When the user is prepared to begin, the timer will start, and the trainer will give verbal instructions for each step of the assembly. If the user has difficulty understanding which parts to pick up and where to place them, the trainer will show a picture from the manual for additional context. Once the assembly is finished, the trainee will complete The Task Load Index Form (Appendix A.2.2), which includes the NASA-TLX questions to assess cognitive load.

Since the trainee used both training methods, the post-experiment procedure is to complete the final questionnaire, which is the Preferred Instruction Medium Form (Appendix A.2.4). Participants are asked about their preference for the training methods, including which method they found more intuitive and detailed in providing information. The remaining questions are open-ended, focusing on their overall experiences with the application and face-to-face training, as well as suggestions for improving both training methods.

In the other scenario, where the assembly is firstly assisted by a human trainer, the user will start by completing the Informed Consent Form, assemble the Lego set, and then fill out the Task Load Index Form. Following this, the user will put on the headset to experience the MR tool training, complete the Tool Usability Form, and finally fill out the Preferred Instruction Medium Form.

7.5 Subjects

Previous studies typically involve around 25 professional participants to evaluate their applications. Given our limited resources, we are unable to recruit real assembly line workers. Instead, we aim to gather students who have some basic experience with head-mounted displays (HMDs) and building Lego sets. These students, although not professional assembly workers, possess the necessary basic skills and familiarity with the technology to provide valuable insights into the usability and effectiveness of our application. This approach allows us to conduct a meaningful evaluation while accommodating our resource limitations.

7.6 Results

In this section, the first chapter details the findings from the pre-experiment questionnaires, outlining participants' demographics, prior experiences, and initial expectations. This is followed by a comprehensive analysis of the quantitative results obtained during the training sessions, including metrics on time efficiency, accuracy, and interaction frequency. Finally, after the training, participants provided detailed feedback on their preferences and experiences, highlighting their views on the intuitiveness and effectiveness of each training method, as well as suggestions for future improvements.

7.6.1 Pre-training results

The user study involved recruiting a total of 16 participants, with 4 trainees assigned to each combination of training methods. Informed consent was obtained from all participants to emphasize the ethical considerations taken to protect participant rights and well-being throughout the research process. The ages of the subjects ranged from 18 to 54 years, with an average age of 26.5 years and a standard deviation of approximately 7.22 years (Appendix A.3 Figure 20). Given that the experiment was conducted at Utrecht University, it was anticipated that the majority of the participants would be Master's students. Consequently, a significant portion of the subjects fell within the 18-24 and 24-34 age groups. This age distribution reflects the typical demographic of Master's students at the university, who are often engaged in advanced studies in fields such as Computer Science. This context is important as it provides insight into the background and experience levels of the participants, which may influence their interactions with the training methods being evaluated.

Unfortunately, the gender distribution among the participants is not as balanced as desired, with 12 male participants and only 4 female participants (Appendix A.3 Figure 21). This imbalance is somewhat expected since most participants are students from Utrecht University, particularly those studying Computer Science, a field traditionally dominated by males. Ideally, a more balanced demographic distribution is required to ensure the reliability and validity of the statistical analysis. A more diverse participant pool would help mitigate any potential biases and provide a more comprehensive understanding of the application's usability and effectiveness across different demographics. Efforts to achieve a more balanced gender distribution in future studies will be crucial for enhancing the generalizability of the findings.

The experiment was conducted in English, however, 14 participants indicated that English is not their primary language (Appendix A.3 Figure 22). Despite this, it is unlikely that a language barrier affected their performance. This is because Utrecht University requires an advanced level of English proficiency for students enrolled in English-taught Master's courses. Consequently, all participants are expected to have a high degree of fluency in English, sufficient to understand and engage with the experimental tasks and materials effectively. This requirement ensures that participants can follow instructions, interact with the MR system, and complete the questionnaires without significant language-related difficulties. Therefore, the results of the experiment are considered reliable and not compromised by language proficiency issues. The consideration is important for maintaining the integrity of the study's findings and ensuring that any conclusions drawn are based on the subjects' interactions with the MR system rather than their language abilities.

Regarding experience with MR, the participants had varying levels of familiarity (Appendix A.3 Figure 23). Six participants reported having no prior experience with MR, seven participants indicated that they had used MR before but not frequently, two participants stated that they use MR regularly, and one participant mentioned using it often. Given the subjects' backgrounds, it was likely that they would have some, albeit limited, experience with MR, as it is a current research topic, and the state-of-the-art courses they have taken require knowledge in this area.

Participants with little to no experience with MR generally performed worse on average with the hardware. They required a longer time to become familiar with the system, particularly with the pinching and poking interactions that are standard for gesture control in all current MR headsets. This initial adjustment period was necessary for these participants to effectively navigate and manipulate objects within the MR environment.

The discrepancy in experience highlights the importance of considering users' familiarity with MR technology when designing and implementing MR systems. It also underlines the need for comprehensive introductory sessions to ensure all participants, regardless of their prior experience, can engage effectively with the MR applications. Addressing these differences in experience is crucial for obtaining accurate and reliable data on the usability and effectiveness of MR systems, as well as for developing training programs that can cater to users with varying levels of expertise. We aimed to mitigate this inconsistency by letting the users take as much time as they need to get comfortable with the application before starting the experiment.

Prior to the experiment, participants were asked to report their susceptibility to motion sickness. Eight participants indicated that they did not suffer from motion sickness, while the other eight mentioned that they occasionally experienced mild motion sickness (Appendix A.3 Figure 24). None of the participants reported suffering from severe motion sickness. Individuals who might have had severe motion sickness were to be excluded from participating in the experiment due to the potential risks of cyber sickness.

Participants who reported mild motion sickness were informed about the concept of cyber sickness and were instructed to stop the experiment immediately if they experienced any symptoms. Given the nature of the experiment, where participants remained seated throughout, the likelihood of experiencing motion sickness was considered low. This precautionary measure ensured the safety and well-being of all participants while allowing the experiment to proceed smoothly. Certifying that participants understood the potential risks and how to manage them was important for maintaining the integrity and ethical standards of the study. The approach helped to minimize any discomfort and ensured that the collected data was not compromised by participants experiencing motion sickness during the experiment.

Finally, participants were asked to rate their experience with building Lego sets (Appendix A.3 Figure 25). The majority (12 participants) indicated that they had built Lego sets before but no longer do so or only rarely engage in this activity. Three participants reported that they build Lego sets regularly, and one participant mentioned that they often build Lego sets.

It was expected that most participants would have some prior experience with building Lego sets. Those who continue to build Lego sets regularly or often performed better on average during the experiment. Their familiarity with the process of sorting pieces and intuitively understanding where each piece should be placed gave them an advantage, allowing them to complete tasks more efficiently even before specific instructions were given. This prior experience with hands-on construction tasks translated into better

performance and quicker adaptation to the training methods being evaluated, highlighting the importance of practical experience in enhancing task performance and emphasizing the potential benefits of recruiting participants with relevant backgrounds for studies involving assembly and construction tasks.

7.6.2 Quantitative results

After completing the initial pre-experiment questionnaire, the trainees were prepared to proceed with the two training methods. None of the subjects reported experiencing motion sickness during or after the MR experiment. This is a positive indicator of the system's usability and comfort, as motion sickness can often be a concern in immersive environments. The entire experiment, from start to finish, took an average of 40 minutes to complete. This time frame included several key components: filling in the pre-experiment questionnaires; the initial UI setup, which familiarized trainees with the system; practicing essential interactions such as poking and pinching, which are critical for navigating and manipulating objects within the MR environment; performing both training methods, which constituted the core activities of the study; and finally, filling out the post-training questionnaires that gathered data on participants' experiences and feedback. The comprehensive nature of the experiment ensured that participants had ample opportunity to engage with the MR system and provide detailed insights into its effectiveness and user-friendliness. This structured approach also helped in systematically capturing the various dimensions of user interaction, from technical setup to experiential feedback.

The quantitative results indicate that performance-wise, assembly training took longer when using a headset, and trainees required approximately 33% more time to complete the experiment using the MR tool. The trainees required an average of 5.07 minutes, with a standard deviation of 1.62 minutes. In contrast, the face-to-face training method took significantly less time, averaging 3.42 minutes with a standard deviation of 0.87 minutes. These results highlight a noticeable difference in the time efficiency of the two training methods. The greater time required for the MR tool can be attributed to the initial learning curve associated with familiarizing participants with the technology and interface. The higher standard deviation in the MR tool's completion time indicates more variability in how quickly different participants adapted to and navigated the tool. Conversely, the face-to-face training method showed less variation in completion times, suggesting a more consistent training pace across participants. This could be due to the direct and immediate feedback provided by a human trainer, which may streamline the learning process. Overall, while the MR tool offers innovative and interactive training opportunities, these findings suggest that it may initially require more time for users to become proficient. However, as users become more accustomed to the MR tool, their efficiency is likely to improve, potentially narrowing the time gap compared to traditional face-to-face training.

The quality of the training outcome was evaluated by counting the number of mistakes made, with fewer mistakes signifying better trainee performance. On average, participants using the MR tool made 0.375 mistakes, translating to an accuracy rate of 98.03%. In comparison, those who underwent face-to-face training made an average of 0.4375 mistakes, corresponding to an accuracy rate of 97.63%. This comparison suggests that, on average, trainees using the MR tool achieved slightly fewer mistakes compared to those using face-to-face training. The lower error rate with the MR tool may be attributed

to its visual and interactive nature, which can provide clearer instructions during the assembly process. In contrast, face-to-face training, while offering direct interaction and communication with a trainer, might involve more subjective interpretations and verbal instructions that could lead to slightly higher error rates.

Across the study, the 16 participants collectively asked a total of 18 questions, split evenly between the two training methods. Specifically, they posed 9 questions to the dialogue agent while using the MR tool and another 9 questions during the face-to-face training sessions.

This equal distribution of questions indicates that participants sought a comparable level of clarification and assistance regardless of the training method. The dialogue agent in the MR tool was as frequently utilized for inquiries as the human trainer in the face-to-face sessions. This suggests that the MR tool’s dialogue agent was effective in engaging participants and encouraging them to seek help when needed, comparable to the human interaction provided in traditional training.

Regarding cognitive load, the average score on the Task Load Index questionnaire for face-to-face training was 12.68 out of 30, with a standard deviation of 1.887. In comparison, the MR training scored slightly higher with an average of 12.93 and a standard deviation of 2.61. These scores indicate that both training methods impose a similar cognitive load on participants. Although the MR training method has a marginally higher cognitive load, the difference is not substantial. The standard deviations suggest a slightly greater variability in cognitive load experiences among participants in the MR training, which could be due to the varying levels of familiarity with the MR interface and technology. The relatively close average scores demonstrate that, despite the added complexity of using MR technology, participants did not experience a significantly higher cognitive load compared to traditional face-to-face training. This suggests that the MR tool, while innovative and technologically advanced, is designed in a way that does not excessively burden users cognitively, making it a viable alternative to conventional training methods.

Given that each participant completed both types of training in different sequences, we need to account for the paired nature of the data when analyzing the differences between the two training methods. This can be effectively addressed using paired statistical tests such as the paired t-test and the Wilcoxon signed-rank test. These tests are designed for paired or dependent samples, meaning the same participants are measured under two different conditions.

To proceed with the analysis, we will first group the data based on the initial training order (i.e., whether the participant started with MR training or face-to-face training). Next, we will use the Shapiro-Wilk test to assess the normality of the data distribution for each dependent variable. If the Shapiro-Wilk test indicates that the data is normally distributed, we will use the paired t-test to compare the means of the two training methods. However, if the Shapiro-Wilk test suggests that the data is not normally distributed, we will employ the Wilcoxon signed-rank test. The Wilcoxon signed-rank test is a non-parametric alternative to the paired t-test and does not assume normality.

	Training Order	Normality	t/W-value	p-value	Significant difference
Performance	MR first, F2F second	No	1	0.01	Yes
	F2F first, MR second	Yes	-3.61	0.008	Yes
Quality	MR first, F2F second	No	3	1	No
	F2F first, MR second	No	2	0.56	No
Questions Asked	MR first, F2F second	No	1.5	0.19	No
	F2F first, MR second	No	7	0.45	No
Cognitive Load	MR first, F2F second	Yes	-1.04	0.32	No
	F2F first, MR second	No	0.101	0.92	No

Table 4. Significance test based on the training order, either MR or face-to-face (F2F) first.

Table 4 presents the results of the paired statistical tests conducted to determine the significance of differences between the MR and face-to-face training methods. Significance is determined by a null hypothesis criteria p lower than 0.05, indicating that the observed differences are statistically significant and not due to random chance.

For the Wilcoxon signed-rank test, the test statistic is denoted by the W -value. This value is derived from the ranks of the differences between paired observations. The W -value is then compared against a critical value from the Wilcoxon signed-rank distribution to determine if the difference is significant. A smaller W -value indicates a more pronounced difference between the pairs.

For the paired t -test, the test statistic is represented by the t -value. The t -value is a ratio that measures the difference between the sample means relative to the variation within the sample data. It is calculated by taking the difference between the means of the paired observations and dividing it by the standard error of the differences. The larger the absolute t -value, the greater the difference between the pairs. This t -value is compared against a critical value from the t -distribution to assess significance.

Analyzing the result table, it is evident that there is a statistically significant difference in performance between the MR training and face-to-face training methods. This difference can primarily be attributed to the additional time required for participants to acclimate to the MR training environment. Specifically, the MR training method demands more initial familiarization, which affects overall performance time.

However, the analysis reveals that for the other three dependent variables (the quality of the training outcomes, number of questions asked, and cognitive load) there is no statistically significant difference between the MR tool and face-to-face training. This indicates that despite the longer time needed to get used to the MR tool, it performs comparably to traditional face-to-face training in terms of the accuracy of the assembly task (as indicated by the number of mistakes made), the frequency of questions asked for clarification, and the cognitive load experienced by the trainees. The lack of significant differences in these areas suggests that the MR tool, once mastered, is just as effective and manageable as face-to-face training. This points to the potential of MR tools to be integrated into training programs to make them more efficient and cost-effective, provided that users are given sufficient time to adapt to the new technology. Future iterations of the MR tool could focus on reducing the learning curve to enhance overall performance and user satisfaction. Given that the analysis of the three dependent variables (quality of training outcomes, number of questions asked, and cognitive load) revealed no statistically significant differences between the MR training and face-to-face training methods, we accept the null hypothesis (H_0) and reject the alternative hypothesis (H_1) for these variables. This means that we do not find sufficient evidence to conclude that there is a difference between the two training methods in terms of these specific outcomes.

The System Usability Scale part of the Tool Usability Form (Figure 13) received an overall score of 74.5, indicating a good usability level for the MR AI-assisted assembly training tool. The score ranges from 0 to 100, with a higher score indicating better usability. A score above 68 is generally considered above average, suggesting that users find the tool relatively easy to use and well-integrated into their tasks.



Figure 13. Individual item results for the System Usability Scale questionnaire.

Examining the individual scores for each question, we see that participants rated the tool highly in terms of ease of use (3.5), integration of functions (4.0), quickness to learn (3.9), confidence in using the tool (3.9), and how frequently they would use it (3.4). These positive ratings highlight the tool’s effectiveness and user-friendly design.

However, there are some areas that need improvement, such as reducing perceived complexity (1.5), the need for support to use the tool (1.6), inconsistencies in the tool (1.8), and awkwardness in use (2.3). Addressing these issues could further enhance the overall user experience and increase the score, potentially making the tool more intuitive and accessible for users.

To evaluate the quality of the user experience, the Tool Usability Form incorporated questions derived from the Technology Acceptance Model questionnaire. The form emphasizes that perceived usefulness and perceived ease of use are critical factors influencing the adoption rate of new technology. Participants rated the perceived usefulness of the MR training tool with an average score of 3.74 out of 5, accompanied by a standard deviation of 0.77. This moderately positive rating suggests that users generally found

the tool beneficial for their training needs. The score reflects a consensus that the MR tool effectively supports the training process, although there is room for improvement to enhance its perceived value further. For perceived ease of use, the MR training tool received an average rating of 3.99 out of 5, with a standard deviation of 0.46. This positive rating indicates that participants found the tool relatively easy to use. The low standard deviation signifies that most users shared a similar positive experience regarding the tool’s usability. This high ease-of-use score is highly important for technology adoption, as it suggests that users can quickly learn and operate the tool with minimal difficulty.

On average, subjects interacted with the menu by pressing buttons 21 times, potentially indicating the level of engagement with the interface. The number of button presses required to complete the assembly is 19, suggesting that some users revisited previous steps to review their work or correct mistakes, as well as ask questions. Additionally, the 3D model was grabbed an average of 18 times during the session. This observation could imply that users found manipulating rotations of the 3D model easier to control than interpreting instructions verbally provided by a trainer. These metrics provide insights into user behavior and preferences during the assembly training. The frequency of menu interactions and 3D model manipulations reflects how participants engaged with the MR tool’s interface and utilized its features.

7.6.3 Post-training results

In the post-training questionnaire, participants were asked about their preferred method of training. Eight participants reported that they preferred the MR tool, finding it to be an effective training method. Only one participant expressed a preference for face-to-face training, while seven participants indicated that they preferred a combination of both the MR tool and face-to-face training methods (Appendix A.3 Figure 26).

The majority’s preference for the MR tool highlights its potential as a valuable training method, possibly due to its interactive and immersive nature. Those who preferred both methods appreciated the complementary strengths of each approach (the immersive, hands-on experience provided by the MR tool and the direct, personalized guidance available through face-to-face training). This mixed-method preference suggests that until dialogue agents fully become human-like, a hybrid approach could offer the most comprehensive training experience, leveraging the advantages of advanced technology while retaining the benefits of traditional, personal interaction.

Regarding which training method felt more intuitive to use, the results were fairly balanced. Five participants indicated that the MR tool was more intuitive, while another five felt that the face-to-face method was more intuitive. Finally, six participants found both methods to be equally intuitive (Appendix A.3 Figure 27).

This distribution suggests that there is no clear consensus on the intuitiveness of either method. The equal preference indicates that both training approaches have their own strengths in terms of user-friendliness and ease of use. Participants who found the MR tool intuitive likely appreciated the 3D model interaction and immersive aspects of the technology, which can make complex tasks easier to understand and perform. On the other hand, those who preferred the face-to-face method might have valued the direct, personal interaction and immediate feedback that this traditional approach offers. The group that found both methods intuitive underscores the potential benefit of integrating both MR and face-to-face training to accommodate different learning preferences and enhance overall training effectiveness.

Finally, when asked which method provided more detailed information, 13 participants indicated a preference for the MR tool, while only 3 participants found the face-to-face method to be more detailed (Appendix A.3 Figure 28).

This overwhelming preference for the MR tool suggests that participants found it to be superior in delivering comprehensive and detailed information. The MR tool's ability to visually demonstrate procedures and provide real-time, interactive feedback likely contributed to this perception. The immersive environment of MR can offer a depth of detail and clarity that is harder to achieve through face-to-face training alone. This descriptive information delivery is crucial for complex assembly tasks where understanding the nuances and specifics is essential for successful execution. The fact that only a small number of participants favored face-to-face training as more detailed indicates that while personal interaction has its benefits, it may not be as effective in conveying intricate details as the MR tool. These findings highlight the potential of MR technology to enhance training programs by providing detailed, easily accessible information, thus improving learning outcomes and efficiency.

For the open-ended questions regarding what participants liked about the MR tool, several key themes emerged. Participants appreciated the visual clarity provided by the MR tool, noting that it is more effective to see things directly rather than having them explained verbally. The MR tool allowed participants to see the different pieces clearly, eliminating any ambiguity in the description of pieces or their placements.

Additionally, the ability to zoom in and rotate the model was highlighted as particularly useful. This feature enabled participants to examine details closely and understand the assembly process from different angles. The tutorial videos were also praised for effectively showcasing the basic input mapping, helping users quickly learn how to interact with the MR environment.

Participants also valued the flexibility of placing the virtual model anywhere and in any orientation. This adaptability was seen as especially beneficial for more complex models, where such features would significantly enhance the assembly process. The precision of the instructions provided by the MR tool was another advantage, as it left no room for confusion about which piece was needed and where it should be placed.

Regarding areas for improvement for the MR tool, participants provided several suggestions. Some participants, particularly those with no prior MR experience, requested additional practice with clicking the buttons. Despite having unlimited time to practice, some trainees chose to rush through this stage and later realized they needed more practice before the assembly process.

Many participants noted that the 'Next Step' button, which was the most frequently used, was difficult to click repeatedly. This issue was exacerbated by the limitations of the headset's hand-tracking capabilities. Although the interface was designed with current standards in mind and works smoothly with controllers, gesture interaction proved challenging due to the suboptimal hand tracking of the available headset. The quality of the camera was another point of criticism. Participants mentioned that better visualization of the real world, where pieces could be more easily recognized, would have prevented some mistakes. In future work, the use of a high-quality headset could mitigate these issues.

Some participants suggested implementing voice activation for the buttons. However, this feature was not included because the current speech-to-text models do not support continuous voice capture. Implementing voice activation would necessitate pressing a button each time a user wants to speak and another button press to stop, effectively doubling the workload compared to the frequent need to press the 'Next Step' button.

These suggestions point to several areas for future improvements. One area is experimenting with headsets that offer enhanced hand-tracking capabilities as well as improved camera quality for better real-world visualization. Additionally, exploring more advanced voice activation technologies that can capture speech continuously without needing additional button presses is recommended.

For the other training method, participants highlighted several advantages of face-to-face training. One notable benefit is that the absence of interface interaction makes communication feel more natural and fluid. Participants appreciated the ability to double-check information with the trainer, making it easier to clarify doubts immediately.

Trainees also mentioned that it feels more comfortable to ask for help in a face-to-face setting, as they receive instant replies to their questions. This immediate feedback is crucial for effective learning and quick problem resolution. Additionally, the opportunity to engage in small talk and discuss topics beyond the task at hand was seen as a positive aspect, fostering a more relaxed and supportive learning environment.

Overall, participants found it easier to make queries to a real person who can provide direct and immediate answers, in contrast to an AI system that requires time to analyze the environment before responding. However, future advancements in AI models could potentially improve the speed and accuracy of responses, making interactions with AI trainers as simple and effective as those with human trainers.

Participants identified several areas for improvement in face-to-face training. They mentioned that ambiguous terms used for pieces or placements often made the process more difficult, and the instructions were not always clear. A visual manual for the assembler would simplify the process, but this is not feasible in current assembly scenarios where manuals contain hundreds of pages and training is conducted verbally.

Trainees noted that visual stimulation is beneficial, and having a visual model similar to the MR tool would enhance understanding. However, in real-life assembly training, where the process is lengthy and involves numerous pieces, it is impractical to have a person provide a 3D representation of the object being assembled due to time and material constraints.

This feedback highlights the need for integrating more visual aids into face-to-face training to improve clarity and reduce ambiguity. Future advancements could focus on developing hybrid training methods that combine the interactive, visual elements of MR tools with the personal interaction of face-to-face training. This approach could provide a more comprehensive and effective training experience, leveraging the strengths of both methods to overcome their respective limitations.

8 Discussion

This project entails the development of a fully functional MR application aimed at assisting with assembly tasks. The study examines the feasibility and technological readiness of creating a standardized method to produce realistic and beneficial augmented graphics in a real, dynamic environment, integrated with a dialogue agent.

The primary goal of the application was to design an effective, efficient, and user-friendly MR system specifically for assembly training. This involved ensuring that the MR tool could deliver high-quality, immersive experiences to facilitate learning and executing assembly tasks. The system is designed to enhance training outcomes by providing intuitive, interactive visual aids seamlessly integrated into the user's physical workspace, while also allowing users the freedom to communicate and ask questions.

This section assesses how well the research questions have been answered, followed by a discussion of the study's limitations, shortcomings, and suggestions for future research.

8.1 Subquestions

From the main research question, three sub-questions related to performance and accessibility were presented. These sub-questions aimed to evaluate the design, effectiveness, efficiency, and user-friendliness of the MR system for assembly training purposes. By addressing these aspects, the study sought to understand how well the MR application could facilitate training and identify any limitations that might impact its broader adoption in practical scenarios.

8.1.1 How can we design an architecture that will take into account the advantages of dialogue agents and MR for assembly?

As detailed in Section 5, we conducted a use case analysis to determine the requirements necessary for creating an easy-to-use and intuitive training application. This involved following design recommendations and guidance from existing software tools used in the industry. All requirements were successfully implemented in the final version of the application, with the system architecture based on previous studies. The architectural layers were structured to be highly modular, adhering to established best practices for MR development. The approach ensures that the application can be easily updated and maintained, facilitating future enhancements and scalability.

8.1.2 How do we integrate dialogue agents with interactive MR systems such that they are compatible and still highly performant?

As outlined in Section 6, we first conducted a feasibility study to determine if the project requirements were achievable. This included a risk analysis to identify potential issues and develop mitigation strategies. Following this, we implemented the proposed system by translating the design and requirements into a functional MR application. The integration of dialogue agents with the MR system was carefully engineered to maintain high performance, ensuring that the interactive elements remained responsive and the user experience was not compromised. This seamless integration was achieved by optimizing the communication between the dialogue agents and the MR system, allowing for real-time interactions and effective training support.

8.1.3 How can we evaluate the efficacy of AI-based MR training?

In accordance with best practices for evaluation described in the related work from Section 2, we employed a combination of questionnaires to measure various aspects of the system. These included the System Usability Scale to assess usability, the NASA Task Load Index to gauge perceived workload, and the Technology Acceptance Model to evaluate acceptance of the technology. Additionally, as seen in Section 7, we analyzed the performance of trainees by examining metrics such as the quality of assembly tasks completed, the number and nature of questions asked, and the frequency and type of user interface interactions. This comprehensive evaluation approach provided a robust assessment of the MR training tool’s effectiveness and highlighted areas for potential improvement.

8.2 Shortcomings and Future Work

The results underscore a pivotal challenge in XR applications: the necessity for communication methods that go beyond mere physical interactions to enable effective human-computer collaboration. For XR systems to reach their full potential, they must incorporate advanced communication channels that facilitate seamless interaction between users and the virtual environment. This could include sophisticated voice recognition systems, natural language processing, and intuitive gesture controls that allow users to communicate effortlessly with the system.

Participants have highlighted the importance of clear and precise instructions, which are sometimes better conveyed through advanced interaction techniques than through traditional communication. For instance, incorporating continuous voice recognition could significantly reduce the need for repetitive button presses, streamlining the user experience. Similarly, improved hand-tracking capabilities can ensure more accurate and responsive gesture controls, making the interaction more fluid and less prone to errors. The latest headsets on the market, such as the Apple Vision Pro, already incorporate state-of-the-art interaction features. For instance, this HMD is capable of accurate eye tracking, which allows users to interact with buttons simply by looking at them and tapping their fingers together to click, providing a more seamless and natural user experience. [3].

The Meta Quest 3, used during the experiment, although a recently released headset, does not perform at the level of more expensive business-to-business products such as the Magic Leap 2 or the Apple Vision Pro. Experimental findings highlighted several limitations of the Meta Quest 3. Participants reported occasional lag, which can disrupt the immersive experience and affect training efficiency. Additionally, the Quest 3’s camera struggled with accurately recognizing black pieces, indicating a need for either a different color scheme for assembly models or the adoption of a higher-quality headset.

Moreover, innovative headsets like Galea integrate cognitive sensors capable of tracking and measuring cognitive load [19]. This advancement can replace traditional user questionnaires, such as the Task Load Index form, with real-time, objective data on user cognitive load. By adopting these advanced HMDs, future experiments can achieve more accurate measurements and provide a more effective and immersive training experience tailored specifically for each user.

Furthermore, integrating advanced communication methods can effectively bridge the gap between the benefits of MR tools and face-to-face training. MR tools provide clear, unambiguous instructions and the ability to manipulate 3D models, while face-to-face training offers the immediacy of human feedback and natural communication ease.

In our study, we used the Wizard of Oz approach to mimic a dialogue agent, rather than training and testing an actual AI model. This allowed us to simulate an interactive experience without fully implementing an AI system. Future work can leverage state-of-the-art models to analyze the potential of AI in enhancing MR applications. For instance, GPT-4o represents a significant advancement toward more natural human-computer interaction. It is capable of processing text, audio, image, and video inputs, and generating corresponding outputs. This model responds to audio inputs in as little as 232 milliseconds, with an average response time of 320 milliseconds, which is comparable to human conversation speed [48].

An important limitation of this study is the composition of the participant pool. The sample predominantly consisted of university students, who may not accurately represent the broader population of potential users, such as professional assembly workers. According to Weibel et al. [77], performing assembly tasks requires cognitive skills such as procedural memory and fine motor skills. While MR can effectively train procedural skills, it is less capable of developing fine motor skills, which are typically acquired through years of experience in the assembly field. Compared to experienced assembly workers, students may exhibit slower performance and lower assembly quality when using the same technology. To address this limitation, future studies should recruit a more diverse participant group, including professional assembly workers, to obtain more generalizable results. This approach will ensure that the findings are applicable to a wider range of users and scenarios. Additionally, the gender imbalance among participants could have influenced the study's outcomes, as different genders might interact with technology in varied ways. Achieving a more balanced gender distribution in future research will help mitigate potential biases and provide a more comprehensive understanding of the MR training system's effectiveness.

Due to resource limitations, we opted to utilize simple assembly tasks, such as assembling Lego models. However, it is important to note that the results obtained from these relatively straightforward tasks may not be entirely reliable when extrapolated to more complex environments, such as an engine assembly line. In more intricate and demanding settings, the challenges and requirements for MR training could be significantly different. Factors such as the precision needed, the variety of components involved, and the potential for human error could all vary substantially compared to assembling Lego models. Conducting a similar study in a complex assembly environment would provide a more comprehensive understanding of how MR applications perform under real-world industrial conditions.

Finally, conducting longitudinal studies, as observed in related research, represents a crucial next step to evaluate both short-term and long-term recall in MR training systems. These studies offer valuable insights into the sustained effectiveness and usability of MR technology over extended periods of time, shedding light on its lasting impacts on learning and performance.

9 Conclusion

This study focuses on the development and evaluation of an MR application designed specifically for training tasks in engine assembly, aiming to quantify its impact on various metrics such as assembly time, quality, questions asked, and cognitive load. By directly comparing AI-assisted MR training with traditional face-to-face training, the research provided valuable insights into the relative strengths and weaknesses of each method.

This comparison helps to understand how modern technology can complement or even surpass traditional training approaches in certain aspects. Our findings indicate that MR training, while requiring more time initially for familiarization with the technology, offers significant advantages such as providing clear, unambiguous instructions and slightly reducing assembly mistakes. In contrast, face-to-face training excels in facilitating natural communication and immediate support but may suffer from ambiguities in instructional delivery.

The feedback from participants underscores the preference for direct interaction with human trainers, who can offer immediate responses and personalized support. In comparison, interactions with AI-driven systems often require processing time to analyze the environment before providing answers. However, advancements in AI technology, exemplified by models like GPT-4o, hold promise for improving response speed and accuracy, potentially bridging the gap between human and AI interaction in training scenarios.

Participants also highlighted the possible benefits of integrating more visual aids into face-to-face training to enhance clarity and reduce ambiguity in instruction. Future developments could explore hybrid training approaches that combine the interactive, visual features of MR tools with the interpersonal dynamics of face-to-face training. This hybrid model could offer a more comprehensive and effective training experience, leveraging the strengths of both methodologies to mitigate their respective limitations.

Moreover, introducing slight automation into training processes could prove beneficial in reducing overall training costs while maintaining or improving training effectiveness. By using a dialogue agent for routine or repetitive training scenarios, resources can be optimized, allowing trainers to focus more on other related tasks.

In conclusion, while this study highlights the current strengths and limitations of MR tools in assembly training, ongoing advancements in both MR technology and AI capabilities offer exciting opportunities for further enhancement. Future research should continue to explore these avenues to refine and optimize MR applications for assembly and other industrial training contexts, ultimately aiming to elevate training efficiency, quality, and learner satisfaction.

References

- [1] P. T. K. Aaron Bangor and J. T. Miller. “An Empirical Evaluation of the System Usability Scale”. In: *International Journal of Human–Computer Interaction* 24.6 (2008), pp. 574–594. DOI: [10.1080/10447310802205776](https://doi.org/10.1080/10447310802205776). eprint: <https://doi.org/10.1080/10447310802205776>. URL: <https://doi.org/10.1080/10447310802205776>.
- [2] M. Aebbersold, T. Voepel-Lewis, L. Cherara, M. Weber, C. Khouri, R. Levine, and A. R. Tait. “Interactive Anatomy-Augmented Virtual Simulation Training”. In: *Clinical Simulation in Nursing* 15 (2018), pp. 34–41. ISSN: 1876-1399. DOI: <https://doi.org/10.1016/j.ecns.2017.09.008>. URL: <https://www.sciencedirect.com/science/article/pii/S1876139917301263>.
- [3] Apple. “Type with the virtual keyboard on Apple Vision Pro”. In: (2024). URL: <https://support.apple.com/guide/apple-vision-pro/type-with-the-virtual-keyboard-tana14220eef/visionos#:~:text=Enter%20text%20with%20the%20virtual%20keyboard&text=Look%20at%20each%20key%2C%20then,show%20special%20characters%20and%20accents..>
- [4] M. R. Bahubalendruni and B. B. Biswal. “A review on assembly sequence generation and its automation”. In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 230.5 (2016), pp. 824–838. DOI: [10.1177/0954406215584633](https://doi.org/10.1177/0954406215584633). eprint: <https://doi.org/10.1177/0954406215584633>. URL: <https://doi.org/10.1177/0954406215584633>.
- [5] M. Banquero, G. Valdeolivas, S. Trincado, N. Garcia, and M.-C. Juan. “Passthrough Mixed Reality With Oculus Quest 2: A Case Study on Learning Piano”. In: *IEEE MultiMedia* 30.2 (2023), pp. 60–69. DOI: [10.1109/MMUL.2022.3232892](https://doi.org/10.1109/MMUL.2022.3232892).
- [6] J. Bentley L & Whitten. *System Analysis & Design for the Global Enterprise*. 7th. McGraw-Hill Irwin, 2007, p. 417.
- [7] S. Borsci, G. Lawson, and S. Broome. “Empirical evidence, evaluation criteria and challenges for the effectiveness of virtual and mixed reality tools for training operators of car service maintenance”. In: *Computers in Industry* 67 (2015), pp. 17–26. ISSN: 0166-3615. DOI: <https://doi.org/10.1016/j.compind.2014.12.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0166361514002073>.
- [8] F. Bosché, M. Abdel-Wahab, and L. Carozza. “Towards a Mixed Reality System for Construction Trade Training”. In: *Journal of Computing in Civil Engineering* 30.2 (2016), p. 04015016. DOI: [10.1061/\(ASCE\)CP.1943-5487.0000479](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000479). eprint: [https://ascelibrary.org/doi/pdf/10.1061/\(ASCE\)CP.1943-5487.0000479](https://ascelibrary.org/doi/pdf/10.1061/(ASCE)CP.1943-5487.0000479). URL: [https://ascelibrary.org/doi/abs/10.1061/\(ASCE\)CP.1943-5487.0000479](https://ascelibrary.org/doi/abs/10.1061/(ASCE)CP.1943-5487.0000479).
- [9] D. A. Bowman. “A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods”. In: *Presence: Teleoperators & Virtual Environments* 11 (2002), pp. 404–424. URL: <https://api.semanticscholar.org/CorpusID:1983971>.
- [10] “Can the Oculus 2 with passthrough API take the place of HoloLens”. In: (2021). URL: <https://www.qualium-systems.com/blog/ar-vr/can-the-oculus-2-with-passthrough-api-take-the-place-of-hololens-2-checking-the-hypothesis/>.

- [11] H. Chen, N. Zendehdel, M. C. Leu, and Z. Yin. “Fine-grained activity classification in assembly based on multi-visual modalities”. In: *Journal of Intelligent Manufacturing* (June 2023). ISSN: 1572-8145. DOI: [10.1007/s10845-023-02152-x](https://doi.org/10.1007/s10845-023-02152-x). URL: <https://doi.org/10.1007/s10845-023-02152-x>.
- [12] P. Cipresso, I. A. C. Giglioli, M. A. Raya, and G. Riva. “The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature”. en. In: *Front. Psychol.* 9 (Nov. 2018), p. 2086.
- [13] L. M. Daling, M. Tenbrock, I. Isenhardt, and S. J. Schlittmeier. “Assemble it like this! - Is AR- or VR-based training an effective alternative to video-based training in manual assembly?” In: *Applied ergonomics* 110 (2023), p. 104021. URL: <https://api.semanticscholar.org/CorpusID:257885935>.
- [14] S. Doolani, C. Wessels, V. Kanal, C. Sevastopoulos, A. Jaiswal, H. R. Nambiappan, and F. Makedon. “A Review of Extended Reality (XR) Technologies for Manufacturing Training”. In: *Technologies* (2020). URL: <https://api.semanticscholar.org/CorpusID:230533679>.
- [15] S. Erol, A. Jäger, P. Hold, K. Ott, and W. Sihm. “Tangible Industry 4.0: A Scenario-Based Approach to Learning for the Future of Production”. In: *Procedia CIRP* 54 (2016), pp. 13–18. URL: <https://api.semanticscholar.org/CorpusID:14203404>.
- [16] S. L. Farra, E. T. Miller, N. Timm, and J. C. Schafer. “Improved Training for Disasters Using 3-D Virtual Reality Simulation”. In: *Western Journal of Nursing Research* 35 (2013), pp. 655–671. URL: <https://api.semanticscholar.org/CorpusID:8390188>.
- [17] C. G. Fidalgo, Y. Yan, H. Cho, M. Sousa, D. Lindlbauer, and J. Jorge. *A Survey on Remote Assistance and Training in Mixed Reality Environments*. 2023. DOI: [10.1109/TVCG.2023.3247081](https://doi.org/10.1109/TVCG.2023.3247081).
- [18] “Fully immersive VR learning solutions for training in hazardous and emergency situations”. In: (2023). URL: <https://flaimsystems.com/>.
- [19] Galea. “The World’s Most Advanced Biosensing Headset”. In: (2024). URL: <https://galea.co/#home>.
- [20] A. Gallagher, E. M. Ritter, H. Champion, G. Higgins, M. Fried, G. Moses, C. Smith, and R. Satava. “Virtual Reality Simulation for the Operating Room: Proficiency-Based Training as a Paradigm Shift in Surgical Skills Training”. In: *Annals of surgery* 241 (Mar. 2005), pp. 364–72. DOI: [10.1002/bjs.1800840237](https://doi.org/10.1002/bjs.1800840237).
- [21] M. Gonzalez-Franco, R. Pizarro, J. Cermeron, K. Li, J. Thorn, W. Hutabarat, A. Tiwari, and P. Bermell-Garcia. “Immersive Mixed Reality for Manufacturing Training”. In: *Frontiers in Robotics and AI* 4 (2017). ISSN: 2296-9144. DOI: [10.3389/frobt.2017.00003](https://doi.org/10.3389/frobt.2017.00003). URL: <https://www.frontiersin.org/articles/10.3389/frobt.2017.00003>.
- [22] J. P. Gownder. “How Enterprise Smart Glasses Will Drive Workforce Enablement”. In: (2016). URL: <https://www.forrester.com/report/How-Enterprise-Smart-Glasses-Will-Drive-Workforce-Enablement/RES133722>.

- [23] T. P. Grantcharov, L. Bardram, P. Funch-Jensen, and J. Rosenberg. “Learning curves and impact of previous operative experience on performance on a virtual reality simulator to test laparoscopic surgical skills”. In: *The American Journal of Surgery* 185.2 (2003), pp. 146–149. ISSN: 0002-9610. DOI: [https://doi.org/10.1016/S0002-9610\(02\)01213-8](https://doi.org/10.1016/S0002-9610(02)01213-8). URL: <https://www.sciencedirect.com/science/article/pii/S0002961002012138>.
- [24] Z. Hu, T. Yu, Y. Zhang, and S. Pan. “Fine-grained Activities Recognition with Coarse-grained Labeled Multi-modal Data”. In: Sept. 2020. DOI: [10.1145/3410530.3414320](https://doi.org/10.1145/3410530.3414320).
- [25] “Introducing Apple Vision Pro: Apple’s first spatial computer”. In: (2023). URL: <https://www.apple.com/newsroom/2023/06/introducing-apple-vision-pro/>.
- [26] R. Kneebone, W. Scott, A. Darzi, and M. Horrocks. “Simulation and clinical practice: Strengthening the relationship”. In: *Medical education* 38 (Nov. 2004), pp. 1095–102. DOI: [10.1111/j.1365-2929.2004.01959.x](https://doi.org/10.1111/j.1365-2929.2004.01959.x).
- [27] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei. “Learning Features and Parts for Fine-Grained Recognition”. In: *2014 22nd International Conference on Pattern Recognition*. 2014, pp. 26–33. DOI: [10.1109/ICPR.2014.15](https://doi.org/10.1109/ICPR.2014.15).
- [28] D. Kulak and E. Guiney. *Use cases: requirements in context*. Addison-Wesley, 2012.
- [29] M. Leap. “Landscape Design”. In: (2019). URL: <https://ml1-developer.magicleap.com/en-us/learn/guides/design-landscape>.
- [30] K. Lee. “Augmented Reality in Education and Training”. In: *TechTrends* 56.2 (Mar. 2012), pp. 13–21. ISSN: 1559-7075. DOI: [10.1007/s11528-012-0559-3](https://doi.org/10.1007/s11528-012-0559-3). URL: <https://doi.org/10.1007/s11528-012-0559-3>.
- [31] A. Li, Z. Lu, L. Wang, T. Xiang, X. Li, and J.-R. Wen. *Zero-Shot Fine-Grained Classification by Deep Feature Learning with Semantics*. 2017. arXiv: [1707.00785 \[cs.CV\]](https://arxiv.org/abs/1707.00785).
- [32] X. Li, W. Yi, H.-L. Chi, X. Wang, and A. P. Chan. “A critical review of virtual and augmented reality (VR/AR) applications in construction safety”. In: *Automation in Construction* 86 (2018), pp. 150–162. ISSN: 0926-5805. DOI: <https://doi.org/10.1016/j.autcon.2017.11.003>. URL: <https://www.sciencedirect.com/science/article/pii/S0926580517309962>.
- [33] A. Liverani, G. Amati, and G. Caligiana. “Interactive control of manufacturing assemblies with Mixed Reality”. In: *Integrated Computer-Aided Engineering* 13 (2006). 2, pp. 163–172. ISSN: 1875-8835. DOI: [10.3233/ICA-2006-13205](https://doi.org/10.3233/ICA-2006-13205). URL: <https://doi.org/10.3233/ICA-2006-13205>.
- [34] M. Lysakowski, K. Zywanowski, A. Banaszczyk, M. R. Nowicki, P. Skrzypczynski, and S. K. Tadeja. *Real-Time Onboard Object Detection for Augmented Reality: Enhancing Head-Mounted Display with YOLOv8*. 2023. arXiv: [2306.03537 \[cs.CV\]](https://arxiv.org/abs/2306.03537).
- [35] N. Macchiarella and D. Vincenzi. “Augmented reality in a learning paradigm for flight aerospace maintenance training”. In: *The 23rd Digital Avionics Systems Conference (IEEE Cat. No.04CH37576)*. Vol. 1. 2004, pp. 5.D.1–5.1. DOI: [10.1109/DASC.2004.1391342](https://doi.org/10.1109/DASC.2004.1391342).

- [36] S. Mann, Y. Yuan, F. Lamberti, A. E. Saddik, R. Thawonmas, and F. G. Prattico. “eXtended meta-uni-omni-Verse (XV): Introduction, Taxonomy, and State-of-the-Art”. In: *IEEE Consumer Electronics Magazine* (2023), pp. 1–9. DOI: [10.1109/MCE.2023.3283728](https://doi.org/10.1109/MCE.2023.3283728).
- [37] N. R. Marc Carrel-Billiard Dan Guenther. “Meeting the new reality: immersive learning”. In: (2021). URL: <https://www.accenture.com/us-en/insights/technology/immersive-learning>.
- [38] B. Martin and B. Hanington. *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Rockport Publishers, 2012.
- [39] K. McMillan, K. Flood, and R. Glaeser. “Virtual reality, augmented reality, mixed reality, and the marine conservation movement”. In: *Aquatic Conservation: Marine and Freshwater Ecosystems* 27 (Sept. 2017), pp. 162–168. DOI: [10.1002/aqc.2820](https://doi.org/10.1002/aqc.2820).
- [40] Microsoft. “Start designing and prototyping”. In: (2022). URL: <https://learn.microsoft.com/en-us/windows/mixed-reality/design/design>.
- [41] “Microsoft is discontinuing Windows Mixed Reality”. In: (2023). URL: <https://www.theverge.com/2023/12/21/24010787/microsoft-windows-mixed-reality-deprecated>.
- [42] P. Milgram and F. Kishino. “A Taxonomy of Mixed Reality Visual Displays”. In: *IEICE Transactions on Information and Systems* 77 (1994), pp. 1321–1329. URL: <https://api.semanticscholar.org/CorpusID:17783728>.
- [43] A. J. Miller and S. Kalafatis. “Mixed Reality Equipment Training: A Pilot Study Exploring the Potential Use of Mixed Reality to Train Users on Technical Equipment”. In: *Proceedings of the 2023 7th International Conference on Virtual and Augmented Reality Simulations*. ICVARS ’23. , Sydney, Australia, Association for Computing Machinery, 2023, pp. 105–113. ISBN: 9781450397469. DOI: [10.1145/3603421.3603436](https://doi.org/10.1145/3603421.3603436). URL: <https://doi.org/10.1145/3603421.3603436>.
- [44] S. Moon, S. Kottur, P. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho, R. Subba, and A. Geramifard. “Situated and Interactive Multimodal Conversations”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by D. Scott, N. Bel, and C. Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1103–1121. DOI: [10.18653/v1/2020.coling-main.96](https://doi.org/10.18653/v1/2020.coling-main.96). URL: <https://aclanthology.org/2020.coling-main.96>.
- [45] C. Moro, J. Birt, Z. Stromberga, C. Phelps, J. Clark, P. Glasziou, and A. M. Scott. “Virtual and Augmented Reality Enhancements to Medical and Science Student Physiology and Anatomy Test Performance: A Systematic Review and Meta-Analysis”. In: *Anatomical Sciences Education* 14.3 (2021), pp. 368–376. DOI: <https://doi.org/10.1002/ase.2049>. eprint: <https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/ase.2049>. URL: <https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/ase.2049>.

- [46] C. Moro, C. Phelps, P. Redmond, and Z. Stromberga. “HoloLens and mobile augmented reality in medical and health science education: A randomised controlled trial”. In: *British Journal of Educational Technology* 52.2 (2021), pp. 680–694. DOI: <https://doi.org/10.1111/bjet.13049>. eprint: <https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13049>. URL: <https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13049>.
- [47] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria. *Recent Advances in Deep Learning Based Dialogue Systems: A Systematic Survey*. 2022. arXiv: [2105.04387](https://arxiv.org/abs/2105.04387) [cs.CL].
- [48] OpenAI. “Hello GPT-4o”. In: (2024). URL: <https://openai.com/index/hello-gpt-4o/>.
- [49] “OpenXR Unifying Reality”. In: (2024). URL: <https://www.khronos.org/openxr/>.
- [50] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur. *TEACH: Task-driven Embodied Agents that Chat*. 2021. arXiv: [2110.00534](https://arxiv.org/abs/2110.00534) [cs.CV].
- [51] R. Palmarini, J. A. Erkoyuncu, R. Roy, and H. Torabmostaedi. “A systematic review of augmented reality applications in maintenance”. In: *Robotics and Computer-Integrated Manufacturing* 49 (2018), pp. 215–228. ISSN: 0736-5845. DOI: <https://doi.org/10.1016/j.rcim.2017.06.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0736584517300686>.
- [52] M. Quest. “Best Practices”. In: (2024). URL: <https://developer.oculus.com/resources/mr-design-guideline/>.
- [53] “Quest 3 vs Quest Pro vs HoloLens 2 (Comparison)”. In: (2023). URL: <https://vr-compare.com/compare?h1=0q3goALzg&h2=-MpSqv-rB&h3=EkSDYv0cW>.
- [54] A. P. Rafael Sacks and R. Barak. “Construction safety training using immersive virtual reality”. In: *Construction Management and Economics* 31.9 (2013), pp. 1005–1017. DOI: [10.1080/01446193.2013.828844](https://doi.org/10.1080/01446193.2013.828844). eprint: <https://doi.org/10.1080/01446193.2013.828844>. URL: <https://doi.org/10.1080/01446193.2013.828844>.
- [55] P. A. Rauschnabel, R. Felix, C. Hinsch, H. Shahab, and F. Alt. “What is XR? Towards a Framework for Augmented and Virtual Reality”. In: *Computers in Human Behavior* 133 (2022), p. 107289. ISSN: 0747-5632. DOI: <https://doi.org/10.1016/j.chb.2022.107289>. URL: <https://www.sciencedirect.com/science/article/pii/S074756322200111X>.
- [56] K. S. B. Robert S. Kennedy Norman E. Lane and M. G. Lilienthal. “Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness”. In: *The International Journal of Aviation Psychology* 3.3 (1993), pp. 203–220. DOI: [10.1207/s15327108ijap0303_3](https://doi.org/10.1207/s15327108ijap0303_3). eprint: https://doi.org/10.1207/s15327108ijap0303_3. URL: https://doi.org/10.1207/s15327108ijap0303_3.
- [57] S. Rokhsaritalemi, A. Sadeghi-Niaraki, and S.-M. Choi. “A Review on Mixed Reality: Current Trends, Challenges and Prospects”. In: *Applied Sciences* (2020). URL: <https://api.semanticscholar.org/CorpusID:212907412>.
- [58] M. Rouse. “What Is Multimodal AI?” In: (2023). URL: <https://www.techopedia.com/definition/multimodal-ai-multimodal-artificial-intelligence>.

- [59] W. T. Ryan Jones. “Mid-Market Technology Trends Report”. In: (2023). URL: <https://www2.deloitte.com/us/en/pages/deloitte-private/articles/technology-trends-middle-market-companies-survey.html>.
- [60] M. V. Sanchez-Vives and M. Slater. “From presence to consciousness through virtual reality”. In: *Nature Reviews Neuroscience* 6.4 (Apr. 2005), pp. 332–339. ISSN: 1471-0048. DOI: [10.1038/nrn1651](https://doi.org/10.1038/nrn1651). URL: <https://doi.org/10.1038/nrn1651>.
- [61] N. Seymour, A. Gallagher, S. Roman, M. O’Brien, V. Bansal, D. Andersen, and R. Satava. “Virtual reality training improves operating room performance: Results of a randomized, double-blinded study”. In: *Annals of surgery* 236 (Oct. 2002), 458–63, discussion 463. DOI: [10.1097/01.SLA.0000028969.51489.B4](https://doi.org/10.1097/01.SLA.0000028969.51489.B4).
- [62] J. Slotwinski and R. Tilove. “Smart assembly: industry needs and challenges”. In: (Aug. 2007), pp. 257–262. DOI: [10.1145/1660877.1660914](https://doi.org/10.1145/1660877.1660914).
- [63] *State-Aware Configuration Detection for Augmented Reality Step-by-Step Tutorials*. 2023. DOI: [10.1109/ISMAR59233.2023.00030](https://doi.org/10.1109/ISMAR59233.2023.00030).
- [64] K. Stanney, R. Mourant, and R. Kennedy. “Human Factors Issues in Virtual Environments: A Review of the Literature”. In: *Presence* 7 (Aug. 1998), pp. 327–351. DOI: [10.1162/105474698565767](https://doi.org/10.1162/105474698565767).
- [65] D. Stefanidis, J. R. Korndorffer, R. Sierra, C. Touchard, J. B. Dunne, and D. J. Scott. “Skill retention following proficiency-based laparoscopic simulator training”. In: *Surgery* 138.2 (2005), pp. 165–170. ISSN: 0039-6060. DOI: <https://doi.org/10.1016/j.surg.2005.06.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0039606005002722>.
- [66] “The 70-20-10 Model”. In: (2016). URL: <https://www.bridgespan.org/insights/the-70-20-10-leadership-development-model>.
- [67] R. Thoppilan et al. “LaMDA: Language Models for Dialog Applications”. In: *CoRR* abs/2201.08239 (2022). arXiv: [2201.08239](https://arxiv.org/abs/2201.08239). URL: <https://arxiv.org/abs/2201.08239>.
- [68] TurboSquid. “3D Lego Bricks”. In: (2018). URL: <https://www.turbosquid.com/3d-models/random-lego-bricks-1313942>.
- [69] Unity. “Sentis Overview”. In: (2023). URL: <https://docs.unity3d.com/Packages/com.unity.sentis@1.5/manual/index.html>.
- [70] Unity. “Mixed Reality Template”. In: (2024). URL: <https://docs.unity3d.com/Packages/com.unity.template.mixed-reality@1.0/manual/index.html>.
- [71] “Virtual Reality”. In: (2024). URL: <https://www.britannica.com/technology/virtual-reality>.
- [72] A. Vodilka, M. Kočiško, S. Konečná, and M. Pollák. “Designing a Workplace in Virtual and Mixed Reality Using the Meta Quest VR Headset”. In: *Advances in Design, Simulation and Manufacturing VI*. Ed. by V. Ivanov, J. Trojanowska, I. Pavlenko, E. Rauch, and J. Pitel. Cham: Springer Nature Switzerland, 2023, pp. 71–80. ISBN: 978-3-031-32767-4.
- [73] M. M. Waisberg Ethan Ong Joshua. “Meta smart glasses-large language models and the future for assistive glasses for individuals with vision impairments”. In: (2023). DOI: [10.1038/s41433-023-02842-z](https://doi.org/10.1038/s41433-023-02842-z). URL: <https://doi.org/10.1038/s41433-023-02842-z>.

- [74] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani. “Explainable Embodied Agents Through Social Cues: A Review”. In: *J. Hum.-Robot Interact.* 10.3 (July 2021). DOI: [10.1145/3457188](https://doi.org/10.1145/3457188). URL: <https://doi.org/10.1145/3457188>.
- [75] C.-Y. Wang, A. Bochkovski, and H.-Y. M. Liao. *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. 2022. arXiv: [2207.02696](https://arxiv.org/abs/2207.02696) [cs.CV].
- [76] Z. Wang, S. Zhang, and X. Bai. “A mixed reality platform for assembly assistance based on gaze interaction in industry”. In: *The International Journal of Advanced Manufacturing Technology* 116.9 (Oct. 2021), pp. 3193–3205. ISSN: 1433-3015. DOI: [10.1007/s00170-021-07624-z](https://doi.org/10.1007/s00170-021-07624-z). URL: <https://doi.org/10.1007/s00170-021-07624-z>.
- [77] S. Weibel, U. Bockholt, and J. Keil. “Design Criteria for AR-Based Training of Maintenance and Assembly Tasks”. In: *Virtual and Mixed Reality - New Trends*. Ed. by R. Shumaker. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 123–132. ISBN: 978-3-642-22021-0.
- [78] X.-S. Wei, Y.-Z. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie. *Fine-Grained Image Analysis with Deep Learning: A Survey*. 2021. arXiv: [2111.06119](https://arxiv.org/abs/2111.06119) [cs.CV].
- [79] S. Werrlich, E. Eichstetter, K. Nitsche, and G. Notni. “An Overview of Evaluations Using Augmented Reality for Assembly Training Tasks”. In: *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 11 (2017), pp. 1080–1086. URL: <https://api.semanticscholar.org/CorpusID:721164>.
- [80] S. Werrlich, P.-A. Nguyen, A.-D. Daniel, C. E. F. Yanez, C. Lorber, and G. Notni. “Design Recommendations for HMD-based Assembly Training Tasks”. In: *SmartObjects@CHI*. 2018. URL: <https://api.semanticscholar.org/CorpusID:19220290>.
- [81] S. Werrlich, P.-A. Nguyen, and G. Notni. “Evaluating the training transfer of Head-Mounted Display based training for assembly tasks”. In: *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. PETRA ’18. Corfu, Greece: Association for Computing Machinery, 2018, pp. 297–302. ISBN: 9781450363907. DOI: [10.1145/3197768.3201564](https://doi.org/10.1145/3197768.3201564). URL: <https://doi.org/10.1145/3197768.3201564>.
- [82] S. Werrlich, K. Nitsche, and G. Notni. “Demand Analysis for an Augmented Reality based Assembly Training”. In: June 2017, pp. 416–422. DOI: [10.1145/3076190](https://doi.org/10.1145/3076190).
- [83] B. G. Witmer and M. J. Singer. “Measuring Presence in Virtual Environments: A Presence Questionnaire”. In: *Presence: Teleoperators and Virtual Environments* 7.3 (June 1998), pp. 225–240. DOI: [10.1162/105474698565686](https://doi.org/10.1162/105474698565686). eprint: <https://direct.mit.edu/pvar/article-pdf/7/3/225/1836425/105474698565686.pdf>. URL: <https://doi.org/10.1162/105474698565686>.
- [84] F. Xu, T. Nguyen, and J. Du. *Augmented Reality for Maintenance Tasks with Chat-GPT for Automated Text-to-Action*. 2023. arXiv: [2307.03351](https://arxiv.org/abs/2307.03351) [cs.HC].
- [85] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang. *Learning to Navigate for Fine-grained Classification*. 2018. arXiv: [1809.00287](https://arxiv.org/abs/1809.00287) [cs.CV].

- [86] U. Zaldivar-Colado, S. Garbaya, P. Tamayo-Serrano, X. Zaldivar-Colado, and P. Blazevic. “A mixed reality for virtual assembly”. In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 2017, pp. 739–744. DOI: [10.1109/ROMAN.2017.8172385](https://doi.org/10.1109/ROMAN.2017.8172385).
- [87] Y. Zang, W. Li, J. Han, K. Zhou, and C. C. Loy. *Contextual Object Detection with Multimodal Large Language Models*. 2023. arXiv: [2305.18279](https://arxiv.org/abs/2305.18279) [[cs.CV](#)].
- [88] W. X. Zhao et al. *A Survey of Large Language Models*. 2023. arXiv: [2303.18223](https://arxiv.org/abs/2303.18223) [[cs.CL](#)].

A Appendix

A.1 Application Enlarged Figures



Figure 14. The start screen of the application which includes the 3D Model to be assembled as well as the onboarding instruction cards.

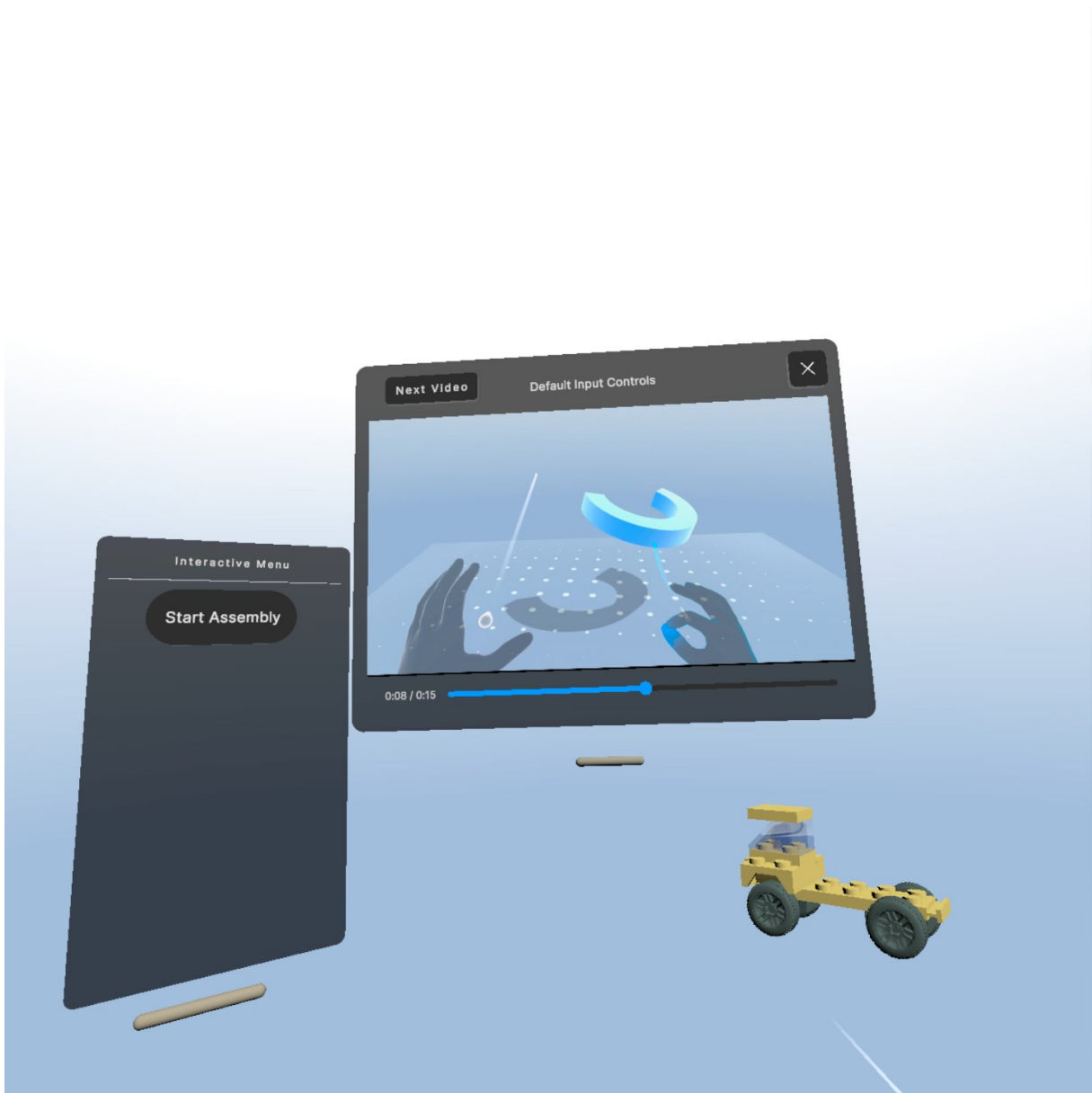


Figure 15. User view after completing the onboarding cards. They will see the interactive menu to their left, the 3D Model to the right, and the input instruction videos towards the middle.



Figure 16. The main UI containing an interactive menu that displays all the interactable buttons, a progress bar as well as a text prompt from the dialogue agent.

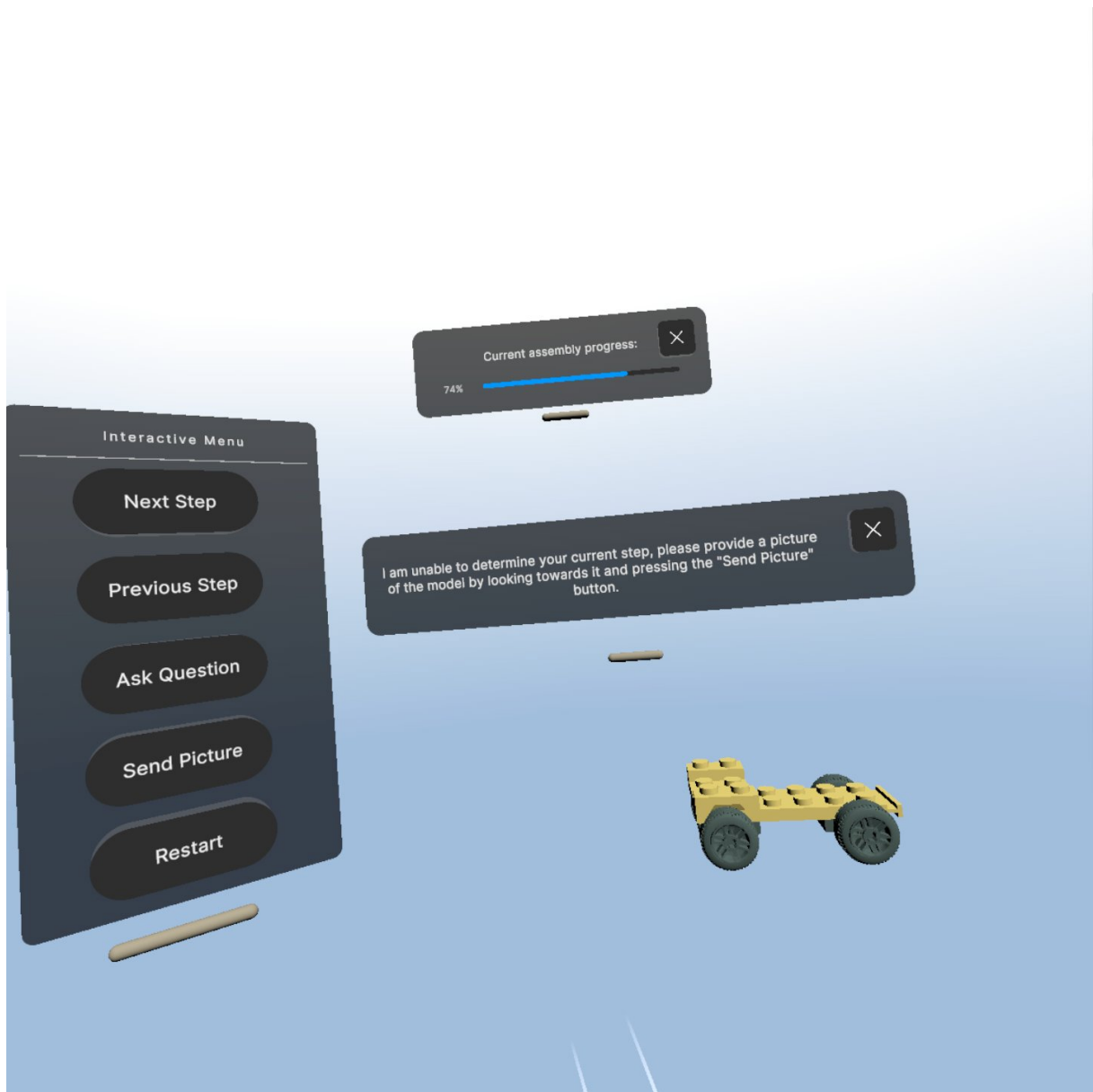


Figure 17. User view after pressing the 'Ask Question' button. In this case, instructions are given within the dialogue agent text box to send a picture of the real world model.



Figure 18. User view after pressing the 'Send Picture' button. The dialogue agent sends a picture from the manual as well as textual assembly instructions.



Figure 19. Additional Helper Menu containing toggles to assist the user in case something goes wrong during the experiment.

A.2 Experiment

A.2.1 Informed Consent Form and Demographics Questionnaire

Informed Consent Form

This consent form will inform you (the participant) about your rights in the upcoming experiment. I (the researcher) have explained the purpose and structure of this experiment, which is part of the Master Thesis of Luca Becheanu, supervised by Wolfgang Hürst.

You are aware that during the experiment, data will be gathered about the interaction between you and the software. This includes basic demographics (gender, age, etc.), interview responses, performance data and screen captures. All data gathered at the experiment may only be for the purpose of this research, including publication in the form of a master thesis. All data gathered in the experiment will be anonymized and treated confidentially.

You understand that participation in the experiment is voluntary: You may abort the experiment at any moment when you desire, and you do not have to provide a reason to us. You are aware that you will suffer no negative consequences from aborting, and that all data gathered during the experiment will be destroyed immediately, and therefore not be used in the research.

You are aware that if you decide to partake in this experiment, it is your responsibility to stop it immediately and inform us in case you experience any discomfort or unwellness, such as dizziness or motion sickness.

You may send further questions about the research to Luca Becheanu (l.becheanu@students.uu.nl) or Wolfgang Hürst (huest@uu.nl). If you suspect that your rights as a participant are violated, you may contact the Research Integrity Committee (vertrouwenspersoon-wi@uu.nl).

1. *

I (the participant) have read the and understood the above text, and I consent that data collected from my participation may be used and published in this research.

Demographics Form

Please fill in exactly one box per question

2. Age *

- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- Above 65
- Would not disclose

3. Gender *

- Man
- Woman
- Prefer not to say

4. Primary Language *

- English
- Other

5. Experience with Mixed Reality (MR) *

- None
- I have used MR before, but not often
- I use MR regularly (more than once per month)
- I use MR often (more than once per week)

6. Motion Sickness *

- I do not suffer from motion sickness
- I occasionally suffer from (mild) motion sickness
- I often suffer from motion sickness, and the symptoms can be severe. (if you select this option, you may not partake in this experiment.)
- Don't know (if you select this option and experience motion sickness, stop and inform us as soon as possible)

7. Experience with building Lego sets *

- None
- I have built Lego sets before, but don't or rarely do it anymore
- I build Lego sets regularly (more than once per month)
- I build Lego sets often (more than once per week)

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

 Microsoft Forms

A.2.2 Tool Usability Form

Tool Usability Form

* Required

Perceived Ease of Use

Give each question a score between one and five based on how much you agree with the question, with 1 being "Strongly Disagree" and 5 being "Strongly Agree"

1. I think I would like to use this tool frequently *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

2. I found the tool unnecessarily complex *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

3. I thought this tool was easy to use *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

4. I think that I would need the support of a technical person to be able to use this tool *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

5. I found the various functions in this tool were well integrated *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

6. I thought there was too much inconsistency in this tool *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

7. I would imagine that most people would learn to use this product very quickly *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

8. I found this tool very awkward to use *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

9. I felt very confident using this tool *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

10. I needed to learn a lot of things before i could get going with this tool *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

Perceived Usefulness

Give each question a score between one and five based on how much you agree with the question, with 1 being "Strongly Disagree" and 5 being "Strongly Agree"

11. Using this tool while assembling would help me complete the tasks faster *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

12. Using this tool would improve my assembling performance *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

13. Using this tool would increase my productivity *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

14. Using this tool enhances my assembling effectiveness *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

15. Using this tool makes it easier to perform assembly tasks *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

16. I find this tool useful for performing assembly tasks *

1	2	3	4	5
---	---	---	---	---

Strongly Disagree

Strongly Agree

Task Load Index

Give each question a score between one and five based on how much you agree with the question, with 1 being "Very Low" and 5 being "Very high"

17. How mentally demanding was the task? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

18. How physically demanding was the task? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

19. How hurried or rushed was the pace of the task? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

20. How successful were you in accomplishing what you were asked to do? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

21. How hard did you have to work to accomplish your level of performance? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

22. How insecure, discouraged, irritated, stressed, and annoyed were you? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

 Microsoft Forms

A.2.3 Task Load Index Form

Task Load Index

Give each question a score between one and five based on how much you agree with the question, with 1 being "Very Low" and 5 being "Very high"

* Required

1. How mentally demanding was the task? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

2. How physically demanding was the task? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

3. How hurried or rushed was the pace of the task? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

4. How successful were you in accomplishing what you were asked to do? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

5. How hard did you have to work to accomplish your level of performance? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

6. How insecure, discouraged, irritated, stressed, and annoyed were you? *

1	2	3	4	5
---	---	---	---	---

Very Low

Very High

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

 Microsoft Forms


A.2.4 Preferred Instruction Medium Form

Preferred Instruction Medium

* Required

1. Which method of training do you prefer? *

- Mixed Reality Tool
- Face-to-face
- Both

2. Which method felt more intuitive to use? * 

- Mixed Reality Tool
- Face-to-face
- Both

3. Which method was more detailed in providing information? *

- Mixed Reality Tool
- Face-to-face
- Both

4. Name one or more things you liked about the Mixed Reality tool *

5. Name one or more things you would like to see improved in the Mixed Reality tool *

6. Name one or more things you liked about the face-to-face training *

7. Name one or more things you would like to see improved in the face-to-face training *

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

 Microsoft Forms

A.3 Results

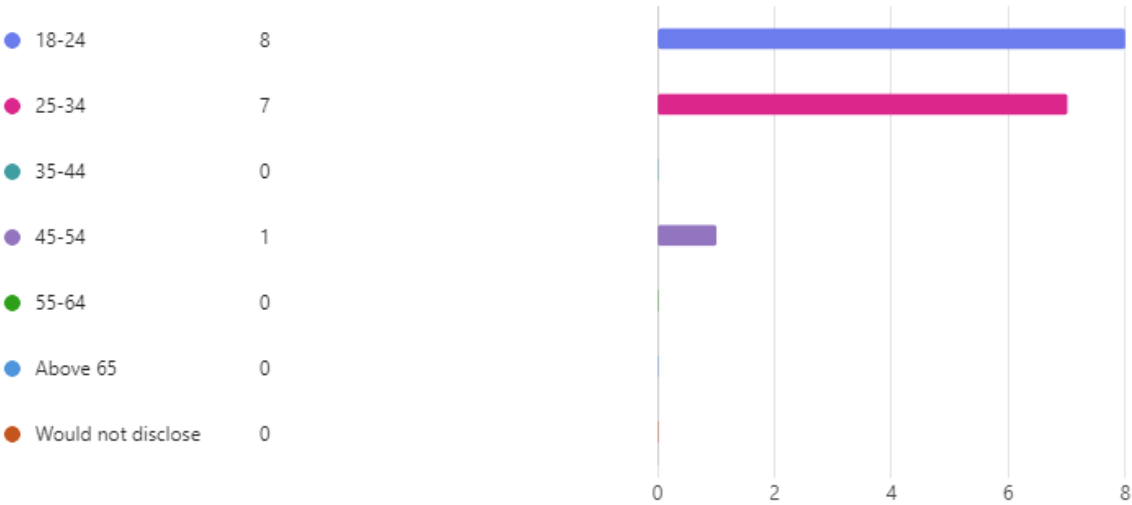


Figure 20. Age distribution of participants.



Figure 21. Gender distribution of participants.



Figure 22. Primary language distribution of participants.



Figure 23. MR experience distribution of participants.

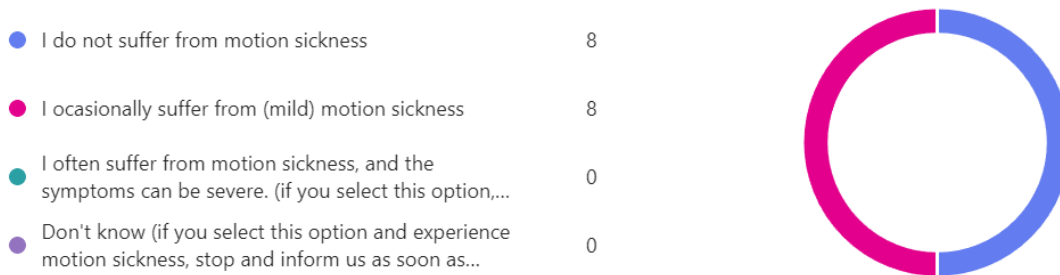


Figure 24. Motion sickness distribution of participants.

Experience with building Lego sets

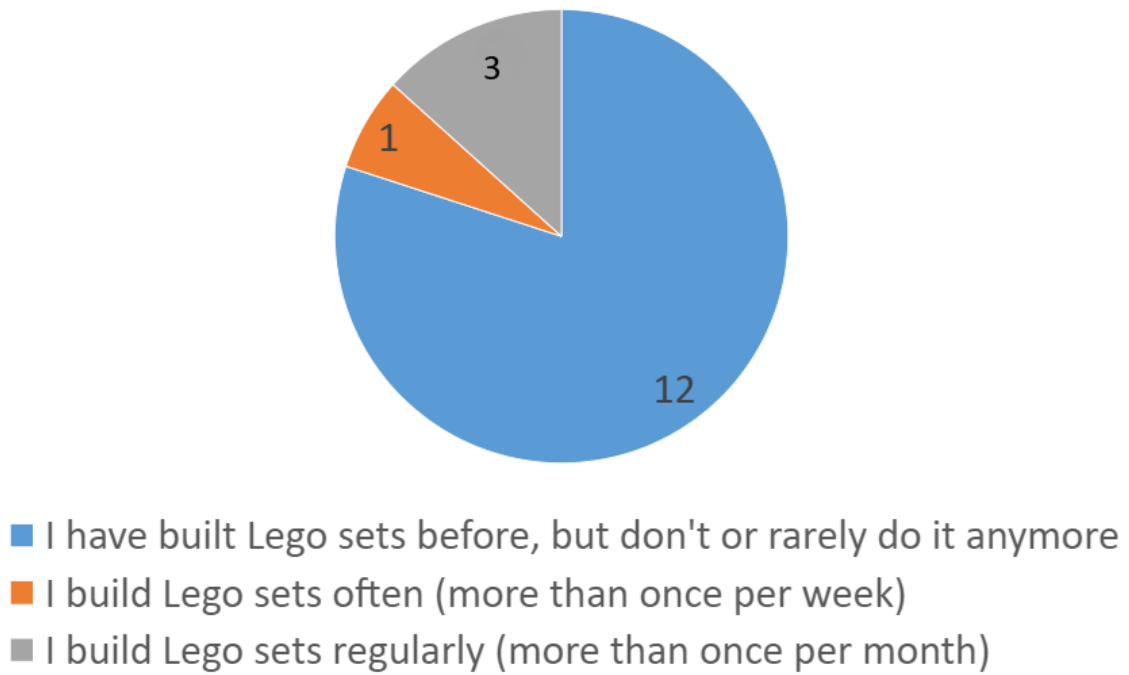


Figure 25. Lego set building experience distribution of participants.

Mixed Reality Tool	8
Face-to-face	1
Both	7



Figure 26. Preferred method of training of participants.

Mixed Reality Tool	5
Face-to-face	5
Both	6



Figure 27. Most intuitive method of training to participants.

● Mixed Reality Tool	13
● Face-to-face	3
● Both	0



Figure 28. Most detailed method of training to participants.