The Impact of Embodiment and Output Modality on Learning with Conversational Agents in VR

Veronika Simmering[1], Thies Pfeiffer[2], and Martin Christof Kindsmüller[3]

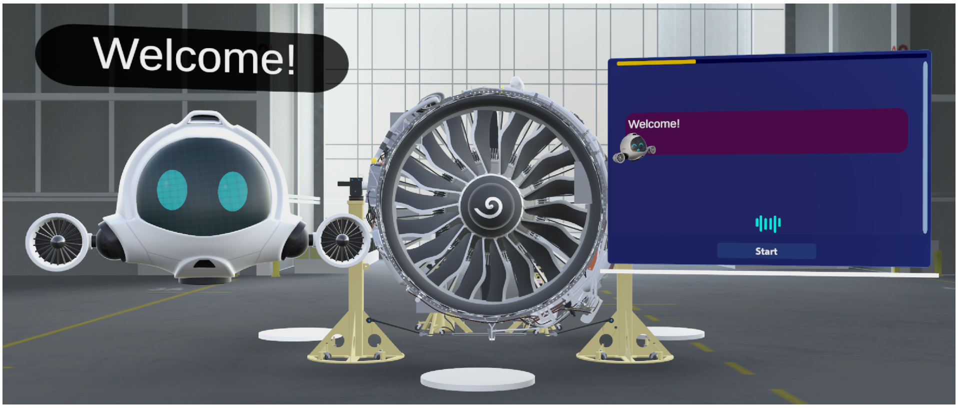

Fig. 1: An overview of the learning application. From left to right: Embodied design variant (Drone); LEAP-1A aircraft engine; non-embodied design variant (Panel).

Abstract: In Virtual Reality (VR) learning applications, the integration of pedagogical agents is particularly promising, as they function as a virtual social support. With the recent developments in Large Language Models (LLMs), conversational language models are now available that enable natural language interaction in a VR environment. We evaluated four possible designs for LLM-based agents ($n = 21$) in an aircraft-engine training. Focusing on the primary questions when bringing LLM-based agents into VR: whether to use text or speech and whether to embody the conversational agent in 3D or not. Pre- and post-tests were used to measure retention and questionnaires were used to measure the User Experience (UX) of different design variants. Contrary to our hypotheses, retention was higher when using a non-embodied design compared to an embodied design. The output modality did not have a significant impact on the learning success, but it did have an impact on the UX.

Keywords: Human-Centered Design, Large Language Models, Virtual Reality

[1] University of Münster, Department of Psychology, Fliednerstraße 21, 48149 Münster, Germany, veronika.simmering@uni-muenster.de, https://orcid.org/0009-0007-0636-5042

[2] University of Applied Sciences Emden/Leer, Faculty of Technology , Constantiaplatz 4, 26723 Emden, Germany, thies.pfeiffer@hs-emden-leer.de, https://orcid.org/0000-0001-6619-749X

[3] Brandenburg University of Applied Sciences, Department of Computer Science and Media, Magdeburger Straße 50 , 14770 Brandenburg an der Havel, Germany, mck@th-brandenburg.de, https://orcid.org/0000-0003-4677-8742

doi:10.18420/vrar2025_12

1 Introduction

Virtual Reality (VR) has many applications, including entertainment, healthcare, engineering, and education [SS16]. In education, VR shows particular promise, with applications ranging from distance learning to training in dangerous situations or the use of expensive or otherwise inaccessible equipment [Ka17], such as aircraft engines. Since aircraft engines are rare, expensive, and difficult to dismantle for training purposes, VR is particularly well suited for training on these devices. In VR, the virtual engine can be viewed and interacted with at actual size without having it physically present in the facilities. As VR learning is usually carried out alone with a Head-Mounted Display (HMD), it could benefit from the addition of pedagogical agents. According to social agency theory, social cues such as embodiment, human voice, and personalized conversational style can motivate learners to invest more effort [Ma14]. The rapid advances in the development of Large Language Models (LLMs) in recent years make it feasible to integrate conversational agents with natural language into VR applications.

Many studies have already focused on pedagogical agents and social agency theory. However, it is unclear to what extent findings from 2D environments can be transferred to VR, given that there are still limited studies on this topic [Da22]. In immersive scenarios in particular, the virtual environment itself may increase the extraneous cognitive load, which could lead to a poorer learning outcome [MP21]. This study therefore investigates how LLM-based agents can be designed for a VR learning environment to enhance learning outcomes while maintaining a positive User Experience (UX).

This raises various questions: Should the LLM-based agent be embodied as a 3D figure or be displayed as a 2D chat window? And should the agent's responses be presented via text only (visual output), or via text and speech (audiovisual output)? Such decisions have a significant impact on the UX, including usability [OB20; Xi23]. Although studies have been conducted on the integration of LLMs in VR applications, such as for learning laboratory tasks [Ko24] and generating dialogue in nursing training [Ka24], the design aspects specifically for immersive learning environments have not yet been widely researched. Steynberg; van Biljon; van der Merwe [SBM24] conducted a relevant study in which they proposed design principles for LLM-based agents based on a literature analysis. However, the authors did not empirically test these principles. Thus, the goal of this research is to create and evaluate a functional prototype integrating LLM-based agents with different design characteristics into a VR learning application.

This research contributes by evaluating different design implementations of LLM-based agents in a VR learning application. Guided by social agency theory, a reversed modality effect and formative interviews we formulated four hypotheses (see Tab. 1): H1 An embodied agent will improve learning compared to a non-embodied agent; H2 An embodied agent will improve UX compared to a non-embodied agent; H3 Presenting information only visually (as text-only) will improve learning compared to presenting information audiovisually

ID	Hypothesis	Rationale
H1	$\text{Retention}_{\text{Embodied}} > \text{Retention}_{\text{Non-embodied}}$	According to the social agency theory, agents with social cues, such as embodiment, can enhance learning [Ma14].
H2	$\text{UX}_{\text{Embodied}} > \text{UX}_{\text{Non-embodied}}$	A thematically fitting embodiment can improve the UX [SBS19].
H3	$\text{Retention}_{\text{visual}} > \text{Retention}_{\text{audiovisual}}$	A reversed modality effect has been found in VR [AS23; Ba20].
H4	$\text{UX}_{\text{Embodied audiovisual}} > \text{UX}_{\text{Embodied visual}}$	Users expect audio modality when using an ECA [Su19].

Tab. 1: Hypotheses

(text + voice) and H4 An embodied agent that speaks will improve the UX compared to a non-speaking agent. The research was conducted in cooperation with Lufthansa Technik.

2 Related Work

Two core aspects of VR are the concepts of immersion and presence. Immersion refers to being immersed in a virtual world, a phenomenon that can be objectively measured by technical features and their ability to influence human perception through various stimuli [DS19; SW97]. Presence describes the subjective experience of perceiving a virtual world as real [Sl03]. This state is favored by a high level of immersion [DS19]. In this study, VR is referred to as immersive VR experienced through an HMD.

2.1 Learning in VR

As described in the introduction, there are many reasons why learning in VR can be useful [Ka17]. VR as a medium brings opportunities for engagement, self-regulation, and motivation. At the same time, however, there are cognitive factors involved in immersive learning that can sometimes hinder learning [MP21]. Information can be presented in various ways and different levels of realism and abstraction in VR. These characteristics present a design challenge: focus on realism and experience, or focus on information access and clarity [Su19].

The HMD used for VR offers the opportunity to experience different modalities: visual, auditory, and haptic with the use of controllers. The modality effect states that presenting complementary audio + visual information can support learning, as information is processed in two separate representational formats within the cognitive system [CP91]. In VR however, there were reversed modality effects observed [AS23; Ba20]. This means visual (image + text) information outperformed audiovisual (image + speech) information, which might be contributed to the highly stimulating environment of a VR application.

2.2 Embodied Conversational Agents

Information in VR can also be transported using an Embodied Conversational Agent (ECA). They are virtual agents that can be used in immersive and non-immersive media. These agents are embodied in their environment, i.e., they have a visual representation with physicality and are not just chatbots, pure voice output, or non-animated objects. In VR, ECAs can increase the social presence [Le06; LSB25], giving the users the sense of talking to another social being. Examples for using ECAs in VR applications are given by Kán; Rumpelnik; Kaufmann [KRK23] or Moore et al. [Mo22] to train socially difficult situations, or by Dai et al. [Da24] to educate students learning to become teachers. In learning scenarios, ECAs appear as pedagogical agents which can provide a higher learning quality and motivation [Ma14] due to the social agency theory, which is described in section 2.3.

There are no clear findings regarding the occurrence of ECAs and their influence. A study by Reinhardt; Hillen; Wolf [RHW20] found that a humanoid agent was favored by their participants. A different conclusion was drawn by Wang; Smith; Ruiz [WSR19], where a miniature version of the agent was favored opposed to one in life-size. A thematically fitting ECA can enhance the UX [SBS19] while Steynberg; van Biljon; van der Merwe [SBM24] recommend using a friendly stylized agent for learning environments in VR.

2.3 Social Agency Theory

A fundamental observation derived from social agency theory is that learners feel socially isolated when learning with multimedia, which reduces their motivational commitment. Using social cues, such as those generated by pedagogical agents, can counteract this. The theory states that learners respond to social cues in learning materials, such as a personal address, natural voice, or visible teacher presence, with a social response. This social perception increases the willingness of learners to actively engage with the learning content, leading to deeper cognitive processing of what has been learned [Ma14]. A personal and conversational style can enhance learning outcomes by mimicking person-to-person communication. Research shows that better learning outcomes could be achieved through dialogue-based learning than with statically presented text [Ma04; MDM03; MM00], which could, for example, be achieved using LLM. However, the data from previous studies is inconsistent. While embodiment can be beneficial, it can also act as a distractor [At02; SAG13]).

2.4 LLMs in XR

LLMs offer the ability to communicate in natural language [Na25]. They can also call functions and provide the appropriate arguments. This makes them a powerful tool in VR, as the presented environment can be manipulated through natural language. In education, LLMs offer many possibilities: they can act as tutors, summarize information, and adapt learning materials [Ba24; Ga23a].

Recent studies have integrated LLMs into VR learning environments. One approach involves the use of ECAs. For instance, Garcia-Pi et al. [Ga23b] integrated LLM-based conversations into a museum application, enabling users to interact with exhibited objects. Compared to a humanoid agent, the participants liked the talking objects the most. Zhu et al. [Zh23] compared the retention from dialogues with an LLM in three different visual representations, showing that the conversation with a humanoid agent led to a higher recall. The dialogue lasted longer than in a conversation with no visual representation or a representation of an object. Another approach to integrate LLMs into XR is provided by Skyba; Pfeiffer [SP24] who used an LLM to enable users in VR to manipulate the scene using natural language, for example, to move their avatar or another object.

3 Formative Analysis

As part of a user-centered design process [DI20] a context analysis and a user analysis were carried out. The user-centered design process focuses on the development of an application around the user and the context to aim for high usability and high UX [DI20]. For that, it must be clear in which context the application is used and what the requirements of the potential users are. These analyses combined with design principles from past studies form the basis for the hypotheses.

3.1 Context Analysis

The LLM-based agent is integrated in an existing learning application. This learning application provides an engine training course for the LEAP-1A aircraft engine (see Fig. 2). The training is intended for apprentices training to become an aircraft engine mechanic. Since aircraft engine models are designed very differently, specific courses are required to learn about their specifics. In this training course, participants will become familiar with the location of the components. Participants will learn basic facts: what the components look like, what they are called, and where they are located. The stakeholders include potential users, apprentices, their instructors, and the application developers.

3.2 Stakeholder interviews

Five stakeholder interviews were conducted to collect the requirements, ideas, and concerns of potential users and other stakeholders. The participants were three aircraft engine apprentices and two VR developers. Despite the initial plans, it was not possible to interview an instructor due to organisational constraints. The interviews were semi-structured to

Fig. 2: 3D model of the LEAP-1A engine

facilitate in-depth questioning. They were conducted in a convergent style, enabling ideas and sketches from previous interviews to be discussed and developed further in later ones. To provide participants with a basis for discussion, the existing version of the engine training application was demonstrated during the interviews. The participants were then interviewed, with the main topics being the appearance and functionality of the LLM-based agent. Each interview resulted in a set of digital post-its and a sketch. All interviews were conducted in German. After analyzing the results, by grouping the notes and comparing the sketches, the main insights from the interviews are the following:

- Three out of five participants wanted an embodied representation; the remaining two wanted a non-embodied one.
- There were heterogeneous ideas about the presentation modality. However, none of the participants wanted an audio-only presentation of the learning information.
- Three roles were identified: tutor, motivator, and organizer.
- The main concern is that the LLM-based agent representing a teacher might provide incorrect information.

These results show that discrepancies are caused by two factors: embodied or non-embodied representation and the modality of the information presented. The identified roles and the sketches provided served as a guide for further development.

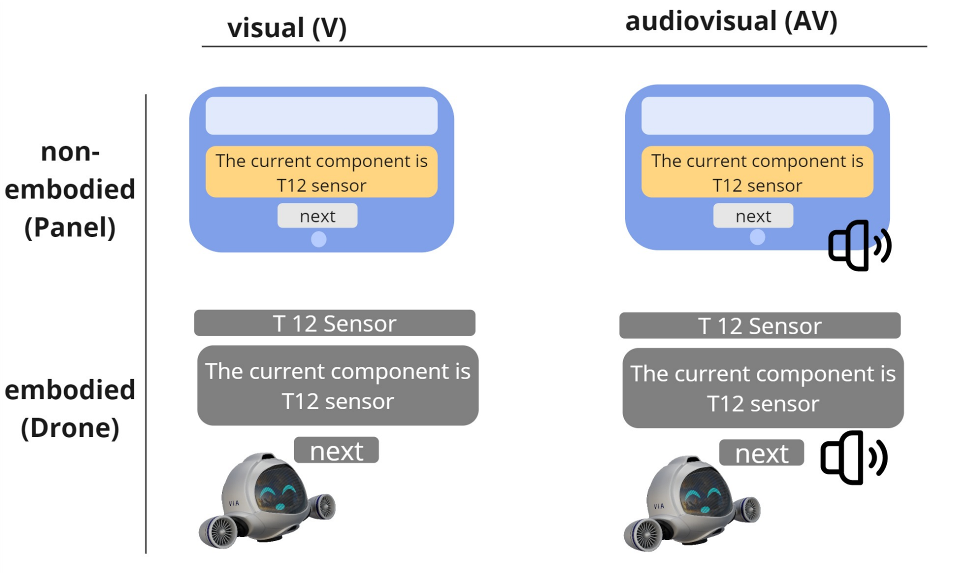
visual (V)
audiovisual (AV)
non-embodied (Panel)
embodied (Drone)
The current component is T12 sensor
next
The current component is T12 sensor
next
T 12 Sensor
The current component is T12 sensor
next
T 12 Sensor
The current component is T12 sensor
next
VIA
VIA

Fig. 3: 2×2 factorial design

4 Prototype and Experimental Design

To test our hypothesis, a fully functional prototype was developed using Unity 3D and the Meta Quest 3. We implemented a 2×2 factorial design, manipulating the LLM-based agent's interface and the output modality (visual vs. audiovisual)(see Fig. 3). Participants were immersed in a virtual aircraft-hangar modeled after those found at Lufthansa Technik sites. An engine model is positioned in the middle. Users can teleport to four teleportation points surrounding it (see Fig. 1), as well as roll the engine to reach each point.

4.1 Interface

One of the manipulated factors is the LLM-based agent's interface. Only the visuals change, while the rest of the scene and the underlying LLM remain unchanged. Specifically, we distinguish between two types of conversational agents:

- An **Embodied**: represented by a drone with expressive animations and spatial presence in the virtual environment (Drone).
- A **Non-embodied**: displayed as a floating chat panel without a physical body in the 3D space (Panel).

The embodied design was selected as a Drone with engines to match the application's theme, as the results by Schmidt; Bruder; Steinicke [SBS19] and Garcia-Pi et al. [Ga23b]

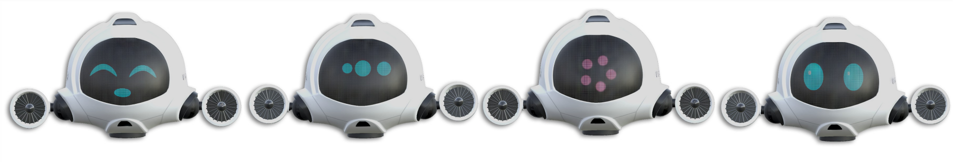

Fig. 4: Drone expressions (left to right: talking, processing, moving, idle)

suggested. The Drone has an idle animation in which it hovers in place. It can either fly next to the current component or be placed freely within the virtual room by the user. It has four different expressions: talking, processing, moving and idle (see Fig. 4) while it always faces the user and maintains eye contact. Text is presented as a speech bubble above.

The non-embodied representation is designed as a floating panel that includes a chat window displaying interactions between the user and the LLM. These interactions are presented alongside a small drone icon, indicating to the user that all responses originate from the same entity. Below the chat window are animations that indicate to the user whether the agent is on standby, processing, or talking. Similar to the Drone, the Panel appears next to the current component and can be placed anywhere on the screen.

4.2 Modality

The second manipulated factor is the agent's output modality. We distinguish between two conditions based on how information is delivered:

- **visual**: The agent's responses are presented solely as written text.
- **audiovisual**: The written text is additionally read aloud using a text-to-speech (TTS) system.

The input remains the same in all conditions: speech via a push-to-talk mechanism.

4.3 Functionality

The three roles identified in the interviews—Tutor, Motivator, and Organizer—were the basis for developing the functionality. The Tutor role presents information, and both the Tutor and Motivator roles actively involve the user. The Organizer role is supposed to help the user to understand the functionality and to navigate the scene. Two modes were designed to facilitate learning: explanation mode and query mode. In explanation mode, information is presented by highlighting components and providing a matching description by the LLM. Users can ask questions about the components or request memory aids. In query mode, the agent highlights a component and asks the user to provide the correct name and description.

In this mode, the user had to be active and speak. When starting the training, a database containing the component names and their official descriptions from the training manual was sent as a string in the first prompt. This method was used because the database was small and to ensure the LLM provided correct information. The initial prompt also stated that the LLM should act as a friendly tutor and only provide the given information without revealing the underlying processes to maintain immersion. GPT-4o provided TTS, STT, and text completion functionality without finetuning. There are four functions that the LLM could call independently:

- save the user's name,
- evaluate the user's answer to a question,
- highlight a specific component,
- show or hide a specific component-group.

5 Method

Quantitative research was conducted with a sample of 21 participants. A within-subject design experiment was conducted, presenting four different combinations to each participant: Drone + AV, Drone + V, Panel + AV, and Panel + V (see Fig. 3).

5.1 Instruments

The dependent variables are UX and retention. UX is measured with the German version of the two seven-point rating scales Efficiency (pragmatic quality) and Stimulation (hedonic quality) from the User Experience Questionnaire (UEQ+ [ST19]). For an additional overall assessment of the application, the scale's Response Quality and Usefulness were selected. Furthermore, questions were asked about the favored condition and about different functionalities of the agent. Retention was measured with a pre- and a post-test. In the pre-test, component descriptions were presented, and the correct component had to be chosen by the participant. Unknown components were collected until they counted eight, so that in each condition two components could be learned. In the post-test, the same procedure in a randomized order was given. All questionnaires as well as pre- and post-tests were presented inside the virtual environment to save time and to not interrupt the experience.

5.2 Participants

The evaluation involved 21 participants (1 female and 20 male). Their ages ranged from 19 to 50 years, with a median age of 27.9 years. The participants were drawn from the mechanical, electronics, and IT training areas, comprising thirteen engine mechanics apprentices, two engine mechanics employees, two electronics apprentices, and four IT apprentices. Participants were recruited through teachers and direct contact at the Lufthansa Technik premises. Participation was voluntary, and participants had varying levels of VR experience.

5.3 Procedure

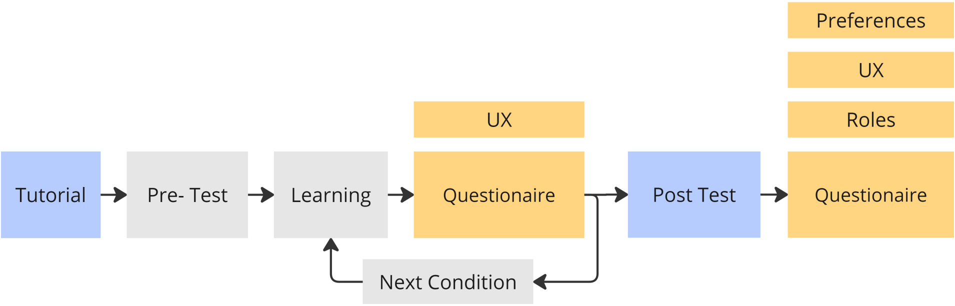

Fig. 5: Procedure for each participant

The evaluation began with a welcome, the collection of demographic data, and the presentation of a consent form. In line with the industry partner's specifications, participants were informed about the study and that all data would be anonymized. The participants were then given the VR headset and started with a tutorial that presented all the main functions. Next, they completed the pre-test, followed by the different learning modes in the four conditions. After that, they were given a UX questionnaire (see Fig. 5). The order of the presented scenarios was counterbalanced in such a way that the interface presentations were grouped, but the order of the different modalities conditions and the order of the interface conditions were randomized. After the learning was completed, the post-test was administered, followed by further questionnaires.

6 Results

The VR sessions took between 20 and 75 minutes with a mean of 37 minutes. Learning success was calculated using the post-test. Each person completed two component tasks in every experimental condition; therefore, for each task, we have a value of either 0 or 1. One person completed only one component in every condition due to incorrect settings. The mean value was then calculated to determine the retention rate. To evaluate the questionnaire scale's Efficiency and Stimulation, ranging from zero to six, a UX-Score was calculated by adding up the score ratings $\mathrm{UX}_i = Efficiency_i + Stimulation_i$ and normalizing them.

$$\hat{\mathrm{UX}}_i = \frac{\mathrm{UX}_i - \min_j(\mathrm{UX}_j)}{\max_j(\mathrm{UX}_j) - \min_j(\mathrm{UX}_j)} \tag{1}$$

To analyze the main effects of Interface and Modality on UX, the four experimental conditions were grouped accordingly.

- **Interface:** Drone (Drone AV and Drone V) and Panel (Panel AV and Panel V)
- **Modality:** Audiovisual (Drone AV and Panel AV) and visual (Drone V and Panel V)

The Figures in the following sections show the mean value (Retention Rate or UX-Score) for each condition, with the width of each bar indicating the value. The standard deviation is also shown.

6.1 Learning

Fig. 6 shows the distribution of the retention rate by Interface and Modality. Contrary to hypothesis H1, it is clear that the retention rate of the participants in the Panel condition (mean 0.64) is higher than in the Drone condition (mean 0.51). In the case of modalities, on the other hand, there are no clear differences. The results of the individual test-conditions are shown in Fig. 7. Drone AV has the lowest mean value with 0.5. Drone V comes second and has a mean value of 0.52. Panel AV and Panel V are equal and have the highest mean score of 0.64. For all conditions, there is a very high standard deviation between 0.37 and 0.42.

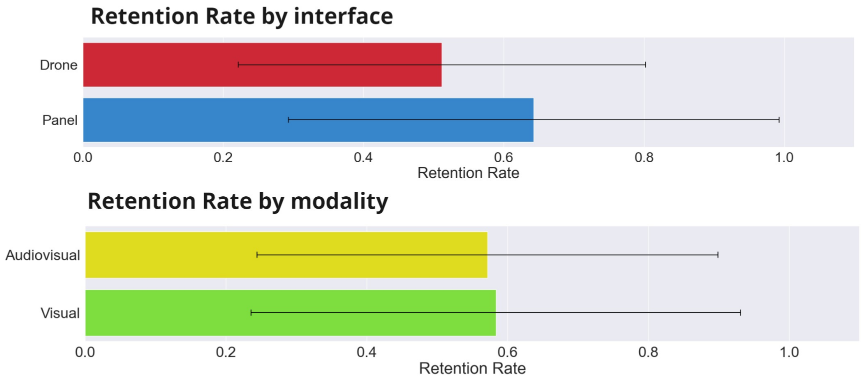

Fig. 6: Retention Rate (0-1) by Interface and Modality

6.2 UX

Fig. 8 shows the UX-Score by interface and modality. The Panel Interface was rated on average at 0.74, which is higher than the 0.70 average rating of the Drone Interface. This finding invalidates hypothesis H2, which predicted a higher UX for the Drone conditions.

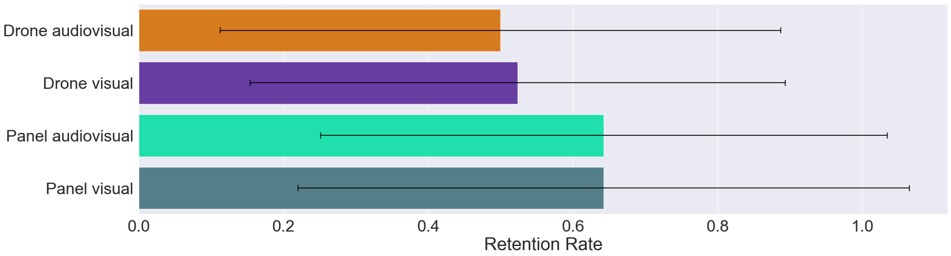

Fig. 7: Retention Rate (0-1) for conditions

The standard deviation of 0.20 for the Drone conditions is greater than that of the Panel conditions, which is 0.13. The audiovisual conditions received a mean rating of 0.77, which was higher than the visual conditions, which received a mean rating of 0.66. Fig. 9 illustrates how the UX scales are distributed across the individual conditions. With a mean value of 0.78, the Drone AV condition has the highest UX-Score. The standard deviation is 0.24. The Panel AV variant follows, with a mean value of 0.77 and a standard deviation of 0.13. The Panel V condition has a mean score of 0.72, while the Drone V condition has a mean score of 0.63 and the highest standard deviation of 0.26. Taking a closer look at the individual scales (see Fig. 10 and Fig. 11) reveals that they were rated differently in terms of Efficiency and Stimulation. In the Efficiency scale, the mean values cluster with a rating around three. Drone AV has the highest mean value rating, at 3.31. The Drone V condition has the largest standard deviation at 0.63. On the Stimulation scale, the differences between the conditions are greater, and the ratings are higher than on the Efficiency scale. Panel AV has the highest mean value of 5.19, while Drone V has the lowest value of 4.05. Once again, conditions involving Drone have higher standard deviations (1.4 and 1.74) than conditions involving UI.

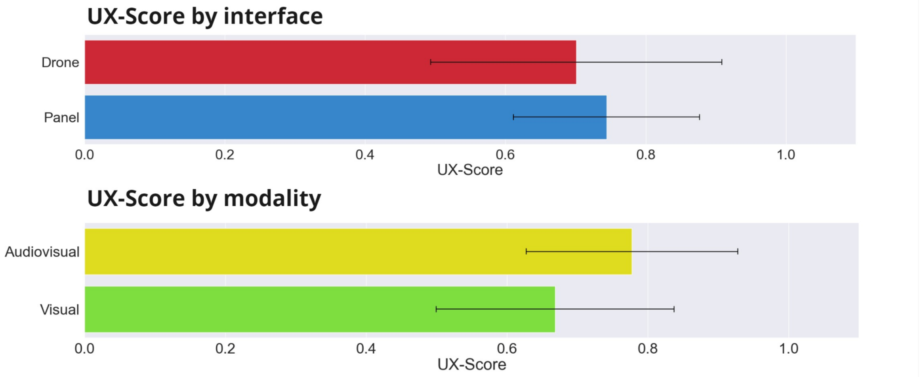

Fig. 8: UX-Scores (0-1) by Interface and Modality

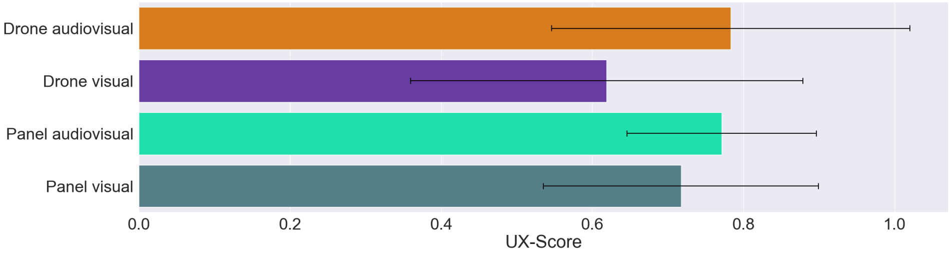

Fig. 9: UX-Scores (0-1) for conditions

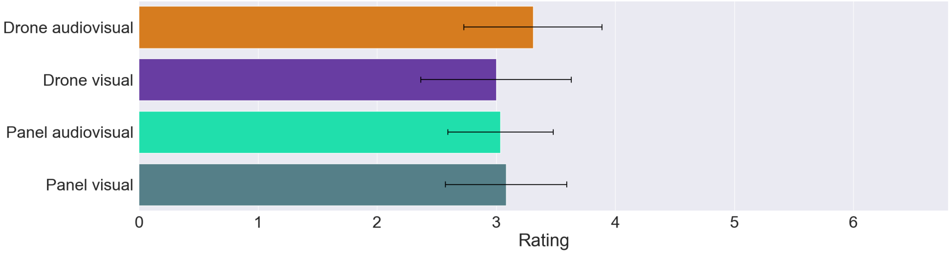

Fig. 10: Efficiency rating (0-6) for conditions

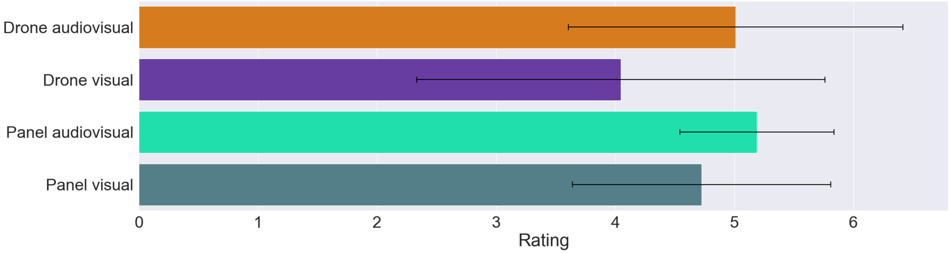

Fig. 11: Stimulation rating (0-6) for conditions

6.3 Testing the hypothesis

A two-way repeated measures ANOVA was conducted for both dependent variables. The results of the hypothesis tests are presented in Tab. 2.

For retention, the analysis revealed a significant main effect of Interface ($F(1, 20) = 4.959$, $p = 0.038$, $\eta_p^2 = 0.199$), but no effect of Modality ($F(1, 20) = 0.026$, $p = 0.874$, $\eta_p^2 = 0.001$). Thus hypothesis H3 could not be supported. Examining the descriptive statistics, retention was higher in the Panel condition than in the Drone condition, indicating that although hypothesis H1 was not supported, the Panel interface exerted a significant positive influence on learning outcomes. Since Retention was measured on an ordinal scale (0, 0.5, 1), the effect of Interface was additionally checked using a Wilcoxon signed-rank test. This test confirmed the ANOVA result, showing that Panel retention scores were significantly higher than Drone scores. No interaction effects were found ($F(1, 20) = 0.033$, $p = 0.858$, $\eta_p^2 = 0.002$).

For UX, the ANOVA showed no significant main effect on Interface ($F(1, 20) = 1.057$, $p = 0.316$, $\eta_p^2 = 0.050$), but a significant main effect of Modality ($F(1, 20) = 14.143$, $p = 0.001$, $\eta_p^2 = 0.414$). This indicates that audiovisual output improved UX compared to visual conditions. Since there were no significant interaction effects found ($F(1, 20) = 1.926$, $p = 0.180$, $\eta_p^2 = 0.088$), hypothesis H4 could not be supported.

Hypothesis	**Dependent Variable**	**Effect Tested**	**p-value (ANOVA)**	**Result**
H1: $\mathrm{Retention}_{\text{Embodied}} > \mathrm{Retention}_{\text{Non-embodied}}$	Retention	Interface	0.038	**not supported**
H2: $\mathrm{UX}_{\text{Embodied}} > \mathrm{UX}_{\text{Non-embodied}}$	UX	Interface	0.316	**not supported**
H3: $\mathrm{Retention}_{\text{visual}} > \mathrm{Retention}_{\text{audiovisual}}$	Retention	Modality	0.874	**not supported**
H4: $\mathrm{UX}_{\text{Embodied audiovisual}} > \mathrm{UX}_{\text{Embodied visual}}$	UX	Interface × Modality	0.180	**not supported**

Tab. 2: Summary of hypotheses and ANOVA results

6.4 Further questions and analysis

As can be seen in Fig. 12, the response quality was rated lower with an average rating of 2.99 (standard deviation 0.25) than the usefulness of the application with an average rating of 5.03 (standard deviation 0.61). 15 out of 21 participants preferred the condition Drone AV. 4 participants preferred the Panel AV condition and one participant preferred the Panel V condition. Furthermore, it was investigated whether the participants' preference for a test condition coincided with the condition in which they actually performed better. In general, only one-third of the participants favored the condition in which they performed best. The remaining participants did not show an interpretable correlation between their preference and higher learning success, that is, although they preferred a certain form of presentation (Drone or Panel), they did not achieve the best results with it. During the experiment, participants were observed to interact with the LLM-based agent in very different ways. While some

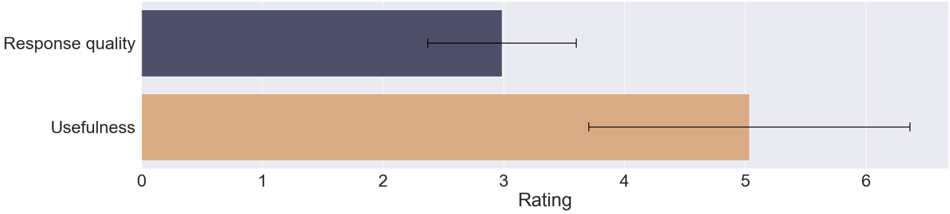

Fig. 12: Ratings (0-6) for Response Quality and Usefulness

answered all questions and asked the LLM-based agent for further information, others avoided speaking, which led to considerable variation in the duration of the experiment. It was also observed that the LLM-based agent did not provide false information, despite this being a concern raised in the formative interviews. Although it had a limited database, it responded to further questions by stating its lack of knowledge.

Furthermore, to complement the students' evaluation, five teachers were interviewed. The participants were first presented with the application and subsequently queried on their understanding of it. All respondents indicated that they would employ an application of this nature in their pedagogical practice.

7 Discussion

The research results were unexpected and opened new perspectives on the questions. The data aligned with the original expectations solely for hypothesis H4. Contrary to expectations, learning outcomes were significantly better with the non-embodied Panel than with the embodied Drone, disproving hypothesis H1. For hypothesis H3, no significant differences were found and no clear trend emerged. Hypothesis H2 was similarly not confirmed. However, an examination of the individual UX scales reveals a discernible trend. Specifically, the Stimulation scale elicited higher ratings for the audio conditions than for the conditions without audio. The Efficiency scale did not reveal significant disparities, suggesting that independent variables exerted minimal influence on the outcome. The presence of latency or other unaccounted factors may have also influenced the Efficiency ratings. Given the lack of support for any of the hypotheses by the data, a thorough examination of the reasons for these outcomes is imperative.

7.1 Interface

No effect could be determined in relation to the social agency theory, as there was no improvement in learning success with the embodied design compared to the non-embodied design. There could be several reasons for this finding. Firstly, the Drone display, with its animations, could have been too distracting. In applications implementing an ECA, users tend to look at it frequently [Lo09; SAG13]. In this case, the balance between motivation and distraction, as described by Moreno [Mo05] and Dehn; Van Mulken [DV00], may not have been optimal. Another explanation could be the level of abstraction and the lack of facial expressions of the embodied design. There may have been too few social cues for the social agency theory to be applicable here. The Drone did not employ gestures or non-verbal communication, nor did it utilize synchronized speaking animations. The aforementioned factors have been shown to be associated with social presence and successful learning [Ma14; PMM21]. Further studies could use eye tracking and cognitive load measurements to investigate test subjects' attention and cognitive load in more detail. It would also be useful to measure social presence and learning motivation using validated questionnaires.

The non-embodied design leveraged its chat window to present additional information. It is posited that the participants' ability to reread the entire conversation may have facilitated a more effective processing of the information. This could have potentially contributed to improved learning outcomes.

Despite the selection of the Drone as the optimal design, no substantial disparities in UX were identified when measured by questionnaires. The discrepancy in the UX-Score is negligible, with the non-embodied Panel representation demonstrating a slight advantage. This could be indicative of the Drone being perceived as more agreeable, though this is only one component of UX. UX is a holistic concept; however, only the Efficiency and Stimulation scales were included in this study. The hypothesis that a preference for the Drone design would be evident in another scale, such as the attractiveness scale, is plausible. Subsequent studies should analyze various aspects of UX in greater detail using additional scales.

7.2 Modality

A clearer difference in UX evaluation was observed for output modalities. Audio output has a particular influence on Stimulation. Opinions on the variant without audio varied greatly. In contrast, no correlation was found between output modality and learning success. No modality effect was found, nor was a reverse modality effect observed, as reported by Albus; Seufert [AS23] and Baceviciute et al. [Ba20]. A salient distinction between the present study and its predecessors pertains to the availability of visual outputs under all conditions. Consequently, participants were capable of comprehending information in the audiovisual condition. This may have counteracted a reverse modality effect. Although the assumption that an embodied design should include audio output, otherwise resulting in a poor UX, could not be empirically confirmed, the data suggests this.

7.3 Limitations

The quantitative study was characterized by a modest sample size of 21 participants, with a notable absence of female representation among the participants, resulting in a group that was imbalanced. In addition, a total of two components were acquired in each condition. This approach can potentially lead to outcomes that are not predictable or consistent. It was reported by some participants that the description, even when considered to be derived from the official training manual, was characterized by ambiguity. Furthermore, some participants reported feelings of discomfort due to the spatial configuration of the study's setting. The influence of these factors on the results is a subject that merits further investigation.

7.4 Design Principles

This led us to develop a set of design principles derived from the findings of the evaluation. The following list contains the items in question:

- **Application-dependent visualization:** The non-embodied design as a floating panel with chat history should be used for learning modes. Embodied visualization should be used for interactive or supporting modes, such as tutorial modes. Ideally, the choice of display should be left to the user.
- **Multimodality in the output:** The use of audio should be optional. It is not decisive for learning success, but it can improve the UX.
- **Information content of the LLM-based agent:** The information content of the LLM-based agent should be increased to enhance its usefulness.
- **Cooperation with trainers:** In order to improve the quality of the learning content, trainers should be involved in its development.
- **Tutorial:** To improve engagement, it is important that all users feel comfortable with the system. Therefore, an extensive tutorial should be provided.

These recommendations can serve as an initial point of reference for designing LLM-based agents for VR learning applications. However, they should be validated and expanded upon in future studies. In particular, there is a need for further investigation into the cognitive effects, as well as the impact on social presence and motivation, of design variants.

7.5 Future research

Future research endeavors present numerous opportunities to further explore this topic. It would be a worthwhile endeavor to experiment with alternative design choices. For instance, one could opt for a humanoid agent, incorporate a wider range of facial expressions, refine the prompts, enhance the autonomy of the agent, or develop an internal model of the user. In terms of methodological decisions, it would be beneficial to measure eye movement, cognitive load, and social presence. Additionally, it would be beneficial to assess the potential impact of the novelty effect on the outcomes of long-term research endeavors. In order to enhance the robustness of the research, it would be advisable to measure design choices individually.

References

[AS23] Albus, P.; Seufert, T.: The Modality Effect Reverses in a Virtual Reality Learning Environment and Influences Cognitive Load. Instructional Science 51 (4), pp. 545–570, 2023, ISSN: 1573-1952, DOI: 10.1007/s11251-022-09611-7.

[At02] Atkinson, R.: Optimizing Learning From Examples Using Animated Pedagogical Agents. Journal of Educational Psychology 94, 2002, DOI: 10.1037/0022-0663.94.2.416.

[Ba20] Baceviciute, S. et al.: Investigating Representation of Text and Audio in Educational VR using Learning Outcomes and EEG. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Acm, Honolulu HI USA, pp. 1–13, 2020, ISBN: 978-1-4503-6708-0, DOI: 10.1145/3313831.3376872.

[Ba24] Bakas, N. P. et al.: Integrating LLMs in Higher Education, Through Interactive Problem Solving and Tutoring: Algorithmic Approach and Use Cases. In (Papadaki, M. et al., eds.): Information Systems. Vol. 501, Springer Nature Switzerland, Cham, pp. 291–307, 2024, ISBN: 978-3-031-56477-2, DOI: 10.1007/978-3-031-56478-9_21.

[CP91] Clark, J. M.; Paivio, A.: Dual coding theory and education. en, Educational Psychology Review 3 (3), pp. 149–210, 1991, ISSN: 1573-336x, DOI: 10.1007/bf01320076.

[Da22] Dai, L. et al.: A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects. Computers & Education 190, p. 104607, 2022, ISSN: 0360-1315, DOI: 10.1016/j.compedu.2022.104607.

[Da24] Dai, C.-P. et al.: Designing Conversational Agents to Support Student Teacher Learning in Virtual Reality Simulation: A Case Study. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. Chi Ea '24, Association for Computing Machinery, Honolulu, HI, USA, 2024, ISBN: 9798400703317, DOI: 10.1145/3613905.3637145.

[DI20] DIN EN ISO 9241-210: Ergonomie Der Mensch-System-Interaktion_- Teil_210: Menschzentrierte Gestaltung Interaktiver Systeme, 2020.

[DS19] Dörner, R.; Steinicke, F.: Wahrnehmungsaspekte von VR. In: Virtual und Augmented Reality (VR/AR): Grundlagen und Methoden der Virtuellen und Augmentierten Realität. Springer, Berlin, Heidelberg, pp. 43–78, 2019, ISBN: 978-3-662-58861-1, DOI: 10.1007/978-3-662-58861-1_2.

[DV00] Dehn, D. M.; Van Mulken, S.: The Impact of Animated Interface Agents: A Review of Empirical Research. International Journal of Human-Computer Studies 52 (1), pp. 1–22, 2000, DOI: 10.1006/ijhc.1999.0325.

[Ga23a] Gan, W. et al.: Large language models in education: Vision and opportunities. In: 2023 IEEE international conference on big data (BigData). Ieee, pp. 4776–4785, 2023, DOI: 10.1109/BigData59044.2023.10386291, accessed: 01/11/2025.

[Ga23b] Garcia-Pi, B. et al.: AllyChat: Developing a VR Conversational AI Agent Using Few-Shot Learning to Support Individuals with Intellectual Disabilities. In (Abdelnour Nocera, J. et al., eds.): Human-Computer Interaction – INTERACT 2023. Vol. 14145, Springer Nature Switzerland, Cham, pp. 402–407, 2023, ISBN: 978-3-031-42292-8, DOI: 10.1007/978-3-031-42293-5_43.

[Ka17] Kavanagh, S. et al.: A Systematic Review of Virtual Reality in Education. Themes in Science and Technology Education 10 (2), pp. 85–119, 2017, DOI: 10.1109/icit58056.2023.10225794.

[Ka24] Kapadia, N. et al.: Evaluation of Large Language Model Generated Dialogues for an AI Based VR Nurse Training Simulator. In (Chen, J. Y. C.; Fragomeni, G., eds.): Virtual, Augmented and Mixed Reality. Vol. 14706, Springer Nature Switzerland, Cham, pp. 200–212, 2024, ISBN: 978-3-031-61040-0, DOI: 10.1007/978-3-031-61041-7_13.

[Ko24] Konenkov, M. et al.: VR-GPT: Visual Language Model for Intelligent Virtual Reality Applications, 2024, DOI: 10.48550/arXiv.2405.11537, eprint: 2405.11537 (cs).

[KRK23] Kán, P.; Rumpelnik, M.; Kaufmann, H.: Embodied Conversational Agents with Situation Awareness for Training in Virtual Reality. The Eurographics Association, 2023, ISBN: 978-3-03868-218-9.

[Le06] Lee, K. M. et al.: Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. International Journal of Human-Computer Studies 64 (10), pp. 962–973, 2006, DOI: 10.1016/j.ijhcs.2006.05.002.

[Lo09] Louwerse, M. M. et al.: Embodied conversational agents as conversational partners. Applied Cognitive Psychology 23 (9), pp. 1244–1255, 2009, DOI: 10.1002/acp.1527.

[LSB25] Lim, S.; Schmälzle, R.; Bente, G.: Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations. Computers in Human Behavior: Artificial Humans, p. 100172, 2025, DOI: 10.1016/j.chbah.2025.100172.

[Ma04] Mayer, R. E. et al.: A Personalization Effect in Multimedia Learning: Students Learn Better When Words Are in Conversational Style Rather than Formal Style. Journal of Educational Psychology 96 (2), p. 389, 2004, DOI: 10.1037/0022-0663.96.2.389.

[Ma14] Mayer, R. E.: Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In: The Cambridge handbook of multimedia learning, 2nd ed. Cambridge handbooks in psychology, Cambridge University Press, New York, NY, US, pp. 345–368, 2014, ISBN: 978-1-107-61031-6, DOI: 10.1017/cbo9781139547369.017.

[MDM03] Mayer, R. E.; Dow, G. T.; Mayer, S.: Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds? Journal of Educational Psychology 95 (4), p. 806, 2003, DOI: 10.1037/0022-0663.95.4.806.

[MM00] Moreno, R.; Mayer, R. E.: Engaging students in active learning: The case for personalized multimedia messages. Journal of Educational Psychology 92 (4), p. 724, 2000, DOI: 10.1037/0022-0663.92.4.724.

[Mo05] Moreno, R.: Multimedia Learning with Animated Pedagogical Agents. In: The Cambridge handbook of multimedia learning. Cambridge Handbooks in Psychology, Cambridge University Press, pp. 507–524, 2005, DOI: 10.1017/cbo9780511816819.032.

[Mo22] Moore, N. et al.: Designing Virtual Reality–Based Conversational Agents to Train Clinicians in Verbal de-Escalation Skills: Exploratory Usability Study. JMIR Serious Games 10 (3), e38669, 2022, DOI: 10.2196/38669.

[MP21] Makransky, G.; Petersen, G. B.: The Cognitive Affective Model of Immersive Learning (CAMIL): A Theoretical Research-Based Model of Learning in Immersive Virtual Reality. Educational Psychology Review 33 (3), pp. 937–958, 2021, ISSN: 1040-726x, 1573-336x, DOI: 10.1007/s10648-020-09586-2.

[Na25] Naveed, H. et al.: A Comprehensive Overview of Large Language Models. ACM Trans. Intell. Syst. Technol. 16 (5), 106:1–106:72, 2025, ISSN: 2157-6904, DOI: 10.1145/3744746.

[OB20] Oprean, D.; Balakrishnan, B.: From Engagement to User Experience: A Theoretical Perspective Towards Immersive Learning. en, Learner and User Experience Research, pp. 199–217, 2020.

[PMM21] Petersen, G. B.; Mottelson, A.; Makransky, G.: Pedagogical Agents in Educational VR: An in the Wild Study. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Chi '21, Association for Computing Machinery, New York, NY, USA, pp. 1–12, 2021, ISBN: 978-1-4503-8096-6, DOI: 10.1145/3411764.3445760, accessed: 01/03/2025.

[RHW20] Reinhardt, J.; Hillen, L.; Wolf, K.: Embedding Conversational Agents into AR: Invisible or with a Realistic Human Body? In: Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction. Acm, Sydney NSW Australia, pp. 299–310, 2020, ISBN: 978-1-4503-6107-1, DOI: 10.1145/3374920.3374956.

[SAG13] Schroeder, N. L.; Adesope, O. O.; Gilbert, R. B.: How Effective Are Pedagogical Agents for Learning? A Meta-Analytic Review. Journal of Educational Computing Research 49 (1), pp. 1–39, 2013, ISSN: 0735-6331, 1541-4140, DOI: 10.2190/EC.49.1.a, accessed: 01/02/2025.

[SBM24] Steynberg, J.; van Biljon, J.; van der Merwe, R.: Design Principles for Pedagogical Agents in a Virtual Reality Learning Environment: Providing Explanations in Real-Time Using Natural Language Processing. In (De Paolis, L. T.; Arpaia, P.; Sacco, M., eds.): Extended Reality. Springer Nature Switzerland, Cham, pp. 227–239, 2024, ISBN: 978-3-031-71713-0, DOI: 10.1007/978-3-031-71713-0_15.

[SBS19] Schmidt, S.; Bruder, G.; Steinicke, F.: Effects of virtual agent and object representation on experiencing exhibited artifacts. Computers & Graphics 83, pp. 1–10, 2019, ISSN: 0097-8493, DOI: 10.1016/j.cag.2019.06.002.

[Sl03] Slater, M.: A Note on Presence Terminology. Presence Connect 3 (3), pp. 1–5, 2003.

[SP24] Skyba, K.; Pfeiffer, T.: Towards natural language understanding for intuitive interactions in XR using large language models, GI VR / AR Workshop, 2024, DOI: 10.18420/vrar2024_0021.

[SS16] Slater, M.; Sanchez-Vives, M. V.: Enhancing Our Lives with Immersive Virtual Reality. Frontiers in Robotics and AI 3, p. 74, 2016, DOI: 10.3389/frobt.2016.00074.

[ST19] Schrepp, M.; Thomaschewski, J.: Handbook for the modular extension of the User Experience Questionnaire, tech. rep., Zugriff am 17.12.2024, PDF report available under, 2019, https://ueqplus.ueq-research.org/Material/UEQ+_Handbook_V6.pdf, accessed: 12/17/2024.

[Su19] Sutcliffe, A. G. et al.: Reflecting on the Design Process for Virtual Reality Applications. International Journal of Human–Computer Interaction 35 (2), pp. 168–179, 2019, ISSN: 1044-7318, DOI: 10.1080/10447318.2018.1443898.

[SW97] Slater, M.; Wilbur, S.: A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. Presence: Teleoperators & Virtual Environments 6 (6), pp. 603–616, 1997.

[WSR19] Wang, I.; Smith, J.; Ruiz, J.: Exploring Virtual Agents for Augmented Reality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Acm, Glasgow Scotland Uk, pp. 1–12, 2019, ISBN: 978-1-4503-5970-2, DOI: 10.1145/3290605.3300511.

[Xi23] Xie, H.: The Applications of Interface Design and User Experience in Virtual Reality. Highlights in Science, Engineering and Technology 44, pp. 189–198, 2023, DOI: 10.54097/hset.v44i.7318.

[Zh23] Zhu, J. et al.: Free-Form Conversation with Human and Symbolic Avatars in Mixed Reality. In: 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Ieee, pp. 751–760, 2023, DOI: 10.1109/ismar59233.2023.00090.