LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-PROC-857890

Towards Anatomy Education with Generative AI-based Virtual Assistants in Immersive Virtual Reality Environments

V. Chheang, R. Marquez-Hernandez, M. Patel, S. Sharmin, D. Rajasekaran, G. Caulfield, B. Kiafar, J. Li, P. Kullu, R. L. Barmaki

December 1, 2023

IEEE International Conference on Artificial Intelligence & extended and Virtual Reality (IEEE AIxVR)
Los Angeles, CA, United States
January 17, 2024 through January 19, 2024

Towards Anatomy Education with Generative AI-based Virtual Assistants in Immersive Virtual Reality Environments

Vuthea Chheang[1]*, Shayla Sharmin[2], Rommy Márquez-Hernández[2], Megha Patel[2], Danush Rajasekaran[2], Gavin Caulfield[2], Behdokht Kiafar[2], Jicheng Li[2], Pinar Kullu[2], Roghayeh Leila Barmaki[2]§
[1]Center of Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, United States
[2]Department of Computer and Information Sciences, University of Delaware, Newark, DE, United States

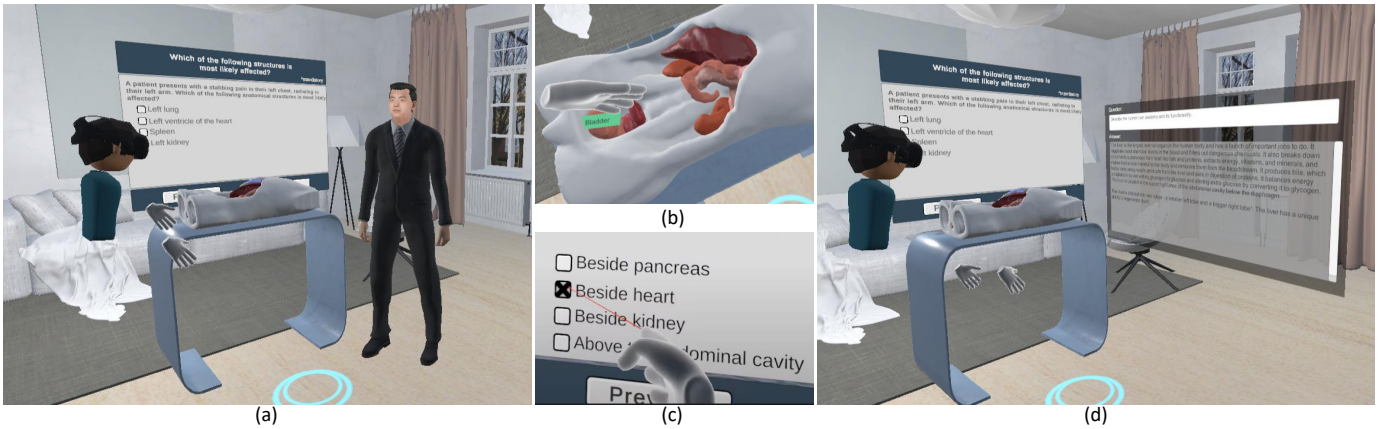

Fig. 1: An immersive VR environment for human anatomy education with generative AI virtual assistants: (a) the user interacts with the virtual assistant as an avatar representation, (b) interactions with abdominal organ 3D models, (c) the user interacts with UI to answer quiz, and (d) configurations of a screen-based virtual assistant.

***Abstract*—Virtual reality (VR) and interactive 3D visualization systems have enhanced educational experiences and environments, particularly in complicated subjects such as anatomy education. VR-based systems surpass the potential limitations of traditional training approaches in facilitating interactive engagement among students. However, research on embodied virtual assistants that leverage generative artificial intelligence (AI) and verbal communication in the anatomy education context is underrepresented. In this work, we introduce a VR environment with a generative AI-embodied virtual assistant to support participants in responding to varying cognitive complexity anatomy questions and enable verbal communication. We assessed the technical efficacy and usability of the proposed environment in a pilot user study with 16 participants. We conducted a within-subject design for virtual assistant configuration (*avatar-* and *screen-based*), with two levels of cognitive complexity (*knowledge-* and *analysis-based*). The results reveal a significant difference in the scores obtained from *knowledge-* and *analysis-based* questions in relation to *avatar* configuration. Moreover, results provide insights into usability, cognitive task load, and the sense of presence in the proposed virtual assistant configurations. Our environment and results of the pilot study offer potential benefits and future research directions beyond medical education, using generative AI and embodied virtual agents as customized virtual conversational assistants.**

***Index Terms*—Generative AI, virtual reality, human-computer interaction, embodied virtual assistants, anatomy education**

I. Introduction

Medical anatomy education, an essential aspect of medical training, necessitates learning the structures and functions of the anatomy in the human body. These skills are vital prerequisites for surgical procedures. Therefore, student awareness of the variation in morphology and the locations of anatomical structures hold significance. Traditionally, medical students learn human anatomy through textbooks, lectures, and dissection of cadavers. However, these approaches have several limitations, such as lack of interactivity, cost, and ethical considerations of cadaveric dissections [1]. Traditional methods of assessing anatomy knowledge encompass a range of approaches, including spotter, written, and oral examinations [2]. For example, anatomy education and assessment have been enhanced by Anderson's modified Bloom's taxonomy, namely Bloom-Anderson principles [3]. The adoption of this taxonomy for anatomy education has a twofold function: first, it considers the cognitive complexity of assignment questions;

*The study was conducted while Vuthea Chheang was a postdoctoral researcher at the University of Delaware. Email: chheang1@llnl.gov.
§Email: rlb@udel.edu.

second, it provides valuable feedback to learners in the context of anatomy education [2], [4].

In recent years, VR has emerged as a viable alternative to traditional anatomy education approaches [5]–[7]. VR enables students to immerse themselves in an engaging and interactive virtual environment where they can interact with 3D anatomy models. In addition, VR allows medical students to conveniently learn via virtual training without worrying about ethical reservations [8], [9]. VR also empowers learners to leverage virtual forums, gamification, peer learning, and virtual laboratories, fostering collaborative and learner-centered environments that align seamlessly with the principles of Bloom's taxonomy [10].

However, most VR-based systems for anatomy education rely on pre-programmed, fixed scenarios that may not adapt well to meet individual learning needs. Here, the support of generative artificial intelligence (AI), such as ChatGPT [11], can be exceptionally advantageous. Compared to conventional virtual assistants, which can be rigid based on predefined rules and templates, generative AI technologies, including chatbots and embodied virtual assistants, have the ability to generate more natural and engaging dialogues that resemble human-human interactions. Although chatbots previously relied on pattern matching and string processing [12], [13], current chatbots use AI, natural language processing (NLP), large-language models (LLMs), and knowledge-based algorithms [14]. These novel technologies empower chatbots to provide more accurate, in-context, personalized, and swift responses to user input while replicating human-human conversations and adapting to the context, levels, and interests of users [15], [16]. Moreover, generative AI leverages large-scale data and information from various sources, including scientific articles, textbooks, and datasets. It generates rich and diverse content to enhance communication and understanding.

This work presents an immersive VR environment designed to support human anatomy education using generative AI conversational assistance (see Fig. 1). We integrated generative AI services (ChatGPT-3.5, OpenAI) into the VR environment, enhanced the embodied virtual avatar representation and realism with lip synchronization, and proposed a new framework to tackle the conversational communication between the user and the virtual assistant using several connected services. The proposed environment has the potential to offer students a more interactive, adaptive, and informative learning experience by offering an embodied virtual assistant. To assess the feasibility of the proposed environment, we developed two different configurations of interaction (*avatar*- and *screen*-based virtual assistants) with two levels of cognitive complexity (*knowledge*- and *analysis*-based) and compared them in a within-subjects pilot user study (n = 16). Our contributions are as follows:

- An immersive VR environment to support the human anatomy education, enabling users to communicate verbally and interact with generative AI-based embodied virtual assistants.
- Results of a pilot user study (n = 16) that provides insights into user performance, usability, task load, and sense of presence in the VR environment.
- An exploratory analysis aimed at identifying potential features, benefits, limitations, and research directions.

II. Related Work

This section provides an overview of previous research on the general use of chatbots for education, with specific focus on VR and virtual assistants in the anatomy education context.

A. Chatbots

Chatbots represent sophisticated computer programs that emulate human-like conversations. They adeptly analyze user inputs and formulate contextually appropriate responses in natural language text. They serve as digital platforms that facilitate concurrent interactions between humans and computers. Chatbots are widespread in many applications, including e-Commerce, education, healthcare, and entertainment [17]. The development of chatbots has become increasingly accessible and versatile. Chatbots such as BERT [18], GPT [19], and Llama [20] have pioneered advancements in NLP, while newer iterations extend their ability in context understanding. These chatbots can assist in the classroom by addressing uncertainties, promoting learning, and providing medical education materials [15].

Instructors can benefit from their help with scheduling, student concerns, and technological support [16]. Chatbots also provide flexible learning help at the convenience of learners, regardless of time or location [14]. In the context of medical education, Termbot [15] offers a convenient way for students to practice medical terminologies. For nursing, AI chatbots are helpful in courses to practice communication, as well as evaluation and intervention skills with patients [16]. Besides, ChatGPT can enable physicians to quickly generate discharge summaries by entering specific facts, concepts, and suggestions [21]. Specifically, ChatGPT has recently been tailored to develop a medical safety LLM framework [22]. This framework involved evaluating ChatGPT's antimicrobial advice across eight hypothetical infection scenarios assessing its suitability, consistency, safety, and stewardship. Finally, within the realm of VR, generative AI with virtual assistants has been used in different applications, including offering support and guidance to individuals with neurological disorders and their caregivers [23], and as an assistive tool for spinal cord surgeries [24].

B. VR-based Anatomy Education Systems

VR and augmented reality (AR) technologies have shown potential in improving medical anatomy education [25]–[29]. Kurt et al. [30] discuss various medical anatomy training approaches, highlighting that cadaver training is restrained by model availability, ethical concerns, and health risks. As an alternative, VR-based training that simulates real-life events sounds appealing. VR-based training can reduce risks, training time, and cost while engaging students.

A survey by Preim [1] emphasizes the importance of learning perspectives, compelling scenarios, and encouraging strategies in anatomy education. Erolin et al. [31] introduced 3D anatomical models for VR anatomy instruction. In their pilot study, most participants found the models easy to interact with and gave positive feedback. Nakai et al. [32] explored the potential of VR-based anatomy courses covering nervous, musculoskeletal, and cardiovascular systems for medical students. They created a VR environment that allows users to manipulate organ anatomy. The findings suggested that the study might provide many 3D models and real-time collaboration.

Kurul et al. [33] studied the impact of immersive VR on anatomy training in physical therapy. Their findings emphasize the value of VR as an alternative training tool. Falah et al. [34] created a VR environment to teach students about medical procedures and anatomical structures. Their solution lets users modify medical data into 3D representations and adjust object sizes in the virtual world. Izard et al. [35] found that VR effectively improves anatomical comprehension in a similar study employing the cranium anatomy. To enhance medical students' learning experience, Saalfield et al. [36] developed a tutoring system that allows teachers to assist students in learning human skull anatomy in a shared virtual environment. Schott et al. [37] proposed a multi-user VR/AR environment for liver anatomy education utilizing clinical examples. According to the study, the prototype could help surgery education in small learning groups and classrooms.

Overall, most of these systems used 3D anatomical models for information visualizations in VR/AR, but they lacked the use of AI or LLMs to make these experiences more learner-centered or self-paced. Our work leverages the power of VR for information visualization by leveraging generative AI and LLMs to offer more genuine interactions with learners.

C. Embodied Virtual Assistants

Anatomy education in VR environments entails a complex and resource-intensive process. It may require a suitable environment and competent instructors to teach human anatomy effectively. The psychological implications of virtual assistants have been a focus of active research [38]–[43]. Kim et al. [44] conducted a study investigating different levels of embodiment in virtual assistants controlled by human-in-the-loop.The study suggested that gesturing and locomotion of the avatar increased trust between the user and the interactive virtual assistant. Haesler et al. [45] conducted a study using *Amazon Alexa*, comparing a voice-only version with an *embodied avatar* version to perform simple everyday tasks. The results showed that the participants preferred the embodied avatar over the voice-only version. In a separate study, Kim et al. [46] analyzed the reduction of task loads with *Amazon Alexa* and findings indicated that using a voice assistant reduced the number of tasks; however, users still expressed uncertainty regarding tasks outside their visual range. On the other hand, the embodied version instilled more confidence in users regarding task completion, thereby fostering greater trust and collaboration between the assistant and the user [47]. These studies collectively indicate that users tend to place more trust in tasks they can visually confirm, highlighting the importance of visual representation in virtual assistants.

Compared to previous work, our VR environment offers a unique advantage by integrating a generative AI-based virtual assistant, *ChatGPT, OpenAI*, as a companion to support learning human anatomy. The embodied virtual assistant enables users to engage in verbal conversations and receive responses to their information queries, resulting in greater confidence and participation. Moreover, it can be used as a source of guidance and to provide users with detailed information. Therefore, the user can seek clarification, ask follow-up questions, and receive personalized explanations to facilitate the understanding of complex medical knowledge as a replacement for a human-in-the-loop assistant or a human trainer.

III. Materials and Methods

The research questions (RQs) to guide our work are the following:

RQ1 How do *configurations* of *avatar*- and *screen*-based virtual assistants influence user performance in anatomy education?

RQ2 To what extent do subjective measures, such as usability, task load, and presence, associate with virtual assistant configurations?

RQ3 What are the advantages, limitations, and potential research directions of using generative AI for anatomy education?

In the following sections, we describe participants, apparatus, study procedure, and study design. Specifically, Fig. 1 to Fig. 3 elaborate more about our pilot user study.

A. Participants

We conducted a priori power analysis to evaluate the sample size with *analysis of variance* (ANOVA) for interaction effects (repeated measures, within factors). Utilizing the *G*Power* statistical analysis software, we calculated the effect size $\eta_p^2 = 0.14$ for two factors, resulting in a sample size of 16 [48].

The study was approved by the University of Delaware's Institution Review Board (#$2136140-1$). The inclusion criteria were participants over the age of 18 with normal to corrected vision, with no known prior history of motion or cyber sickness. Out of the 20 registrants, four participants were excluded due to vision (one individual), motion sickness (two individuals), and VR discomfort (one individual). Thus, 16 participants from the University of Delaware were successfully recruited in our study. The demographic information of the participants are provided in Table I.

B. Apparatus

Fig. 2 shows an overview of the system architecture for the proposed VR environment designed for human anatomy education, featuring generative AI virtual assistants. The VR environment was developed using *Unity* game engine (version $2019.4.34f1$). Customization of 3D models, including the living room, and the incorporation of additional models from

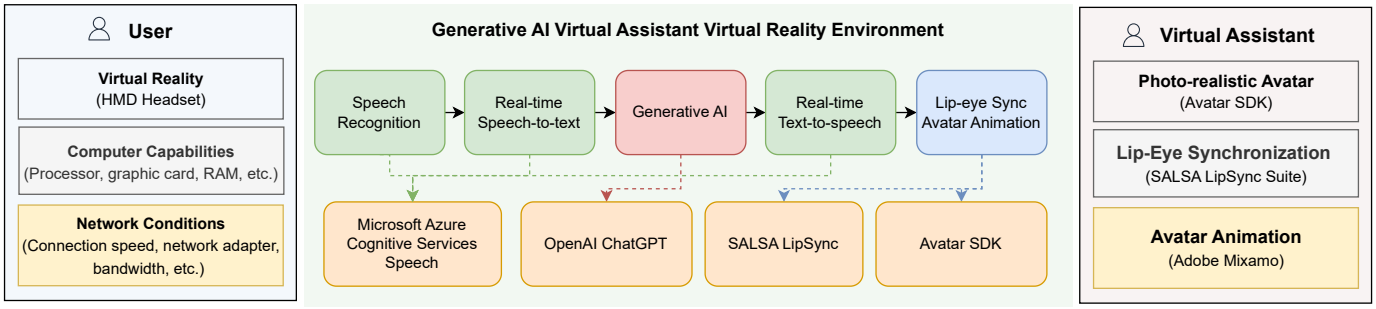

Fig. 2: System architecture of our proposed VR environment with the generative AI-based embodied virtual assistant.

TABLE I: Participant background and characteristics ($n = 16$).

Characteristics	Value	Mean
Age	[20 – 35]	26 ± 4.97
Gender		
Male	6	(37.50%)
Female	10	(62.50%)
Education		
Bachelor's program	6	(37.50%)
Master's program	3	(18.75%)
Doctoral program	7	(43.75%)
Medical anatomy knowledge		
Not much	5	(12.50%)
Basic	11	(87.50%)
Health centered classes		
No	8	(50.00%)
Yes	8	(50.00%)
VR experience		
Never used before	6	(37.50%)
Used a few times	3	(18.75%)
Used several times	5	(31.25%)
Regular use	2	(12.50%)
Handedness		
Left	2	(12.50%)
Right	14	(87.50%)

Sketchfab, alongside *OpenHELP* organ models, were integral to the VR environment's development [49]. Participants were asked to navigate the VR system and engage with the 3D model by grasping, resizing, and rotating them to understand their functionalities during the training session. We used the *Valve Index* VR headset, controllers, lighthouses, and its components within the VR setup.

To enhance user interactions within the VR environment, we utilized the capabilities of the *Virtual Reality Toolkit* (VRTK). This toolkit enabled us to implement fundamental interactions such as teleportation, object manipulation, and interactions with the user interface (UI). Moreover, we used the VR questionnaire toolkit [50] to develop the UI for the VR quiz. We integrated *ChatGPT* (OpenAI, USA) to provide services as an intelligent conversational agent for answering questions. We also used an AI-based library (*Avatar SDK*, Itseez3D Inc., USA) to create a photo-realistic model for the virtual assistant's avatar presented in the user study. This virtual avatar was animated to provide gestures with facial expressions, further enhancing user engagement. We leveraged the Microsoft Azure Speech service to enable natural interactions, utilizing text-to-embodiment capabilities for text-to-speech and speech-to-text. Text-to-embodiment refers to the conversion of text responses from generative AI into the virtual avatar's voice and facial expressions and vice versa, e.g., participant's speech to text.

C. Study Procedure

Fig. 3 presents an overview of the study procedure. To accommodate potential learning curves, we counterbalanced the order of the conditions and the level of cognitive complexity. We walked each participant through the study's procedure. Ensuring their full understanding of the consent form was a priority before requesting their signature. The participants were given the opportunity to become familiarized with the VR headset and controllers and learn how to interact with the virtual assistant in the training environment. They received guidance on asking questions to the virtual assistant by pressing a button on the controller while posing their query, then releasing it to allow the system to process and respond. Additionally, they learned the distinctions in responses between the avatar and the text screen. The study commenced upon participants' confidence on the technology.

To start the study, participants found themselves in a virtual living room environment with a virtual cadaver, a screen displaying a question with multiple-choice answers, and their initial virtual assistant. Interacting within this environment, they could teleport around the room, move the screens around, and manipulate the organs within the cadaver. When ready, participants selected the "Next" button to receive their first question. They were informed they could ask the virtual assistant as many questions as they wished to arrive at the answer to the multiple-choice question, with no pressure to provide a correct response. The only condition was that they could not ask the virtual assistant the entire question directly.

After obtaining an answer, participants selected it from the multiple-choice options and clicked the "Submit" button. Subsequently, they were presented with another question featuring a different cognitive level within the same virtual assistant configuration. Upon completing tasks with their first virtual assistant configuration, participants took a break and completed a short mid-questionnaire on a computer. Following

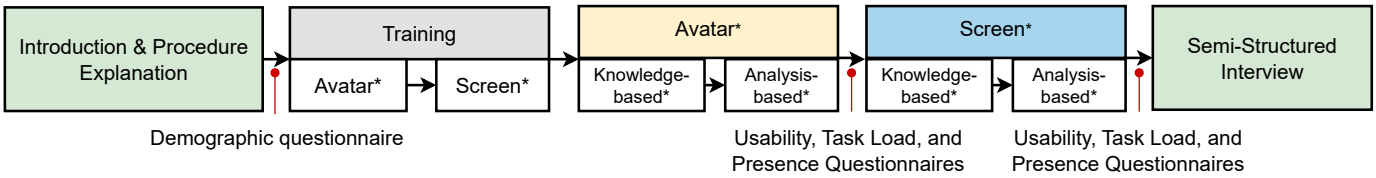

Fig. 3: Overview of the study procedure. The order of the conditions (marked with *) was counterbalanced.

this, they repeated the process of answering two questions with varying levels of cognitive complexity using the other virtual assistant configuration. Once this phase concluded, participants were again given a break to complete a post-questionnaire. Lastly, they had the opportunity to ask any questions and provide qualitative feedback.

D. Study Design

Our study was designed as a 2×2 within-subject experiment (2 *configurations* $\times$ 2 *levels of cognitive complexity*). Each participant was randomly assigned to start with either the *avatar* or the *screen* configuration during the study. They were then presented with two questions of varying cognitive complexity levels for each virtual assistant configuration. We ensured that the opportunity to initiate with either the avatar or the screen was evenly distributed among all participants. This random assignment and counterbalancing helped mitigate potential bias resulting from a learning effect.

1) ***Independent Variables****:* In our study, we defined the virtual assistant configuration and difficulty level as independent variables.

a) ***Configuration****:* Within the virtual environment, we introduced two types of generative AI virtual assistants: *avatar* and *screen* (see also Fig. 1).

- *Avatar*: an embodied avatar equipped with audio and lip synchronization, seamlessly integrated with *Microsoft Azure* text-to-speech and generative AI services to respond to questions.
- *Screen*: a screen displayed text responses generated by generative AI service alongside the participant's question.

b) ***Level of Cognitive Complexity****:* We had four questions categorized in two sets.

- *Knowledge-based*: These questions required no analytical thinking or in-depth understanding and fell under the foundational or first level of Bloom's taxonomy, namely "knowledge" or "remembering".
- *Analysis-based*: These questions demanded more in-depth analysis, corresponding to the fourth level of Bloom's taxonomy, namely "analysis".

These four multiple-choice, scenario-based questions focus on diagnosing medical conditions based on specific symptoms and involved anatomical structures. It includes how different symptoms, such as chest pain, difficulty breathing, abdominal pain, and neurological problems, can be linked to specific organs or systems of the human body, such as the heart, lungs, digestive organs, and nervous system. The goal is to determine the most likely affected area or organ responsible for the presented symptoms. Each configuration contains one knowledge-based and one analysis-based questions. The configuration and the order of questions were counterbalanced.

2) ***Dependent Variables****:* Throughout the study, we recorded the participant's selected answer, task completion time, and the number of interactions with the virtual assistant as the dependent variables. All this data was automatically logged into a *CSV* file for further analysis.

- *Task Completion Time*: the duration between the participant posing the question and submitting the answer.
- *Number of Interactions*: the number of times the participant requests information from the virtual assistant.
- *Score*: a variable indicating whether their answer to the question was correct or incorrect.

3) ***Questionnaires****:* As part of our evaluation, we gathered not only performance data but also valuable insights into participants' subjective experiences through the administration of standardized questionnaires. These questionnaires were designed and administered using the *Qualtrics* survey platform. The following are the specific dimensions we assessed:

- *Usability*: We assessed usability using the System Usability Scale (SUS) questionnaire [51], which comprises ten questions with a 5-point Likert-scale from "*strongly disagree*" to "*strongly agree*". The final SUS score was calculated on a scale from 0 to 100 (0 – 50: not acceptable, 51 – 67: poor, 68: okay, 69 – 80: good, 81 – 100: excellent) [52].
- *Task Load*: To gauge the subjective task load experienced by participants, we employed the NASA Task Load Index (NASA TLX) questionnaire [53]. This questionnaire consists of six questions assessing mental demand, physical demand, temporal demand, performance, effort, and frustration.
- *Presence*: We also evaluated the sense of presence within the virtual environment using *igroup Presence Questionnaire* (IPQ) [54], [55]. This questionnaire has 14 questions categorized as general presence, spatial presence, involvement, and experienced realism. Responses are recorded on a 7-point Likert scale, ranging from "*strongly disagree*" to "*strongly agree*".

4) ***Semi-structured Interviews****:* After the completion of all previous tasks, we solicited qualitative feedback from participants through semi-structured interviews. Participants were asked the following questions:

- What is your feedback on the VR environment and configurations of the virtual assistant?

TABLE II: Summary of descriptive results of user performance.

Variable	Task Completion Time (s)	Number of Interactions	Scores
Avatar	159.23 (149.25) [26.38]	5.96 (11.87) [2.10]	0.59 (0.49) [0.08]
Knowledge-based	117.74 (72.61) [18.15]	5.31 (11.22) [2.80]	0.75 (0.44) [0.11]
Analysis-based	200.72 (192.59) [48.14]	6.62 (12.83) [3.20]	0.43 (0.51) [0.12]
Screen	159.06 (152.94) [27.03]	3.65 (2.74) [0.48]	0.50 (0.50) [0.09]
Knowledge-based	177.50 (200.57) [50.14]	4.31 (3.36) [0.84]	0.56 (0.51) [0.12]
Analysis-based	140.62 (85.96) [21.49]	3.00 (1.82) [0.45]	0.43 (0.51) [0.12]

All entities are in the following format: mean value (standard deviation) [standard error].

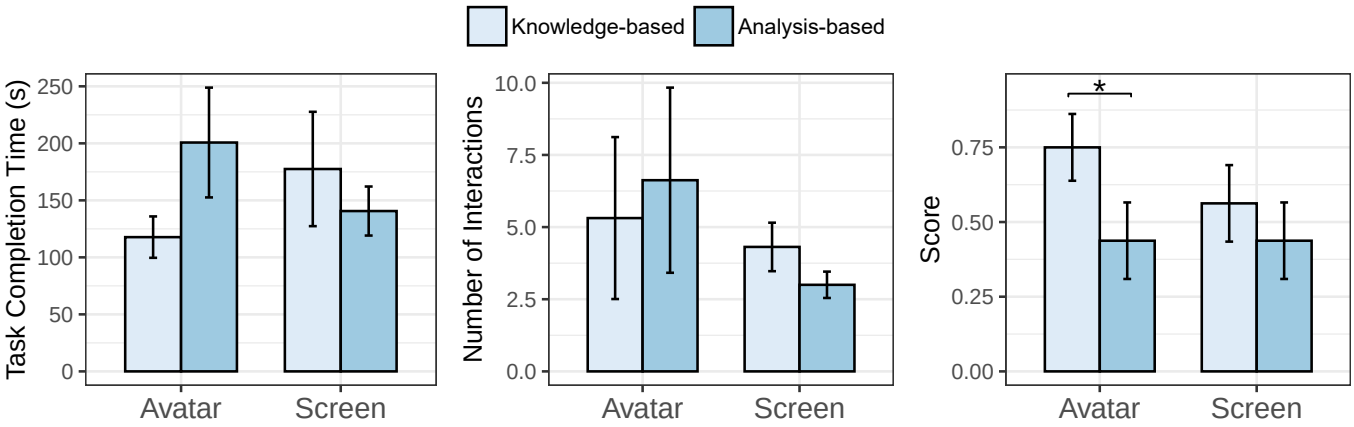

Fig. 4: Results of user performance (n = 16), including (left) task completion time, (middle) number of interactions with the virtual assistant, and (right) score.

- Do you have any questions or suggestions?

IV. Results

In the following sections, we present the results for both descriptive and statistical analysis of user performance, questionnaire outcomes, and general feedback.

A. User Performance Results (RQ1)

We used *RStudio* with the R programming language to conduct a thorough statistical analysis. Our selected method, a two-way ANOVA for dependent variables, enabled us to examine the variables in depth. As we explore our analysis further, a summary of descriptive results related to objective user performance measures is shown in both tabular (Table II) and graphical (Fig. 4) formats.

1) Task Completion Time (TCT): We found no significant differences between the *configuration*, *level of cognitive complexity* and their interaction effect on task completion time. On average, participants responded faster to *knowledge-based* questions in the *avatar* configuration. In the *screen* configuration, *knowledge-based* questions were responded more slowly. The results show that participants spent more time interacting with the *avatar* when tackling *analysis-based* questions, while the *screen* configuration led to more extended task durations for *knowledge-based* questions. Descriptive results for total completion time, however, showed only a minimal difference between the *avatar* (M = 159.23 s, SD = 149.25) and *screen* (M = 159.06 s, SD = 152.94) configurations. It indicates that both configurations are comparable to assist the user in solving the tasks.

2) Number of Interactions: Descriptive results show that all tasks in the *avatar* configuration require more interactions and requests with the virtual assistant, particularly in the context of *analysis-based* questions. The results could indicate that users need more explanation through verbal communication with the virtual avatar compared to the display screen.

3) Score: On average, participants scored higher in solving questions within the *avatar* configuration than in the *screen* configuration. Regarding the *level of cognitive complexity*, there was a significant difference between configurations ($F(1, 15) = 4.62$, $\mathbf{p = 0.046}$, $\eta_p^2 = 0.07$). Subsequent pairwise t-tests indicated a statistically significant difference between *knowledge-based* and *analysis-based* questions (t = -1.78, df = 15, $\mathbf{p = 0.04}$) within the *avatar* configuration (see Fig. 4). This finding shows that participants who engaged with *knowledge-based* questions in the *avatar* configuration obtained higher scores and completed tasks more swiftly compared to *analysis-based* questions. This finding aligns with the Bloom's taxonomy about the level of cognitive complexity as well, because *knowledge-based* questions are less cognitively demanding.

B. Questionnaire Results (RQ2)

In the following sections, we present the subjective results obtained from our questionnaires, offering insights into usability, task load, and presence within our virtual environment.

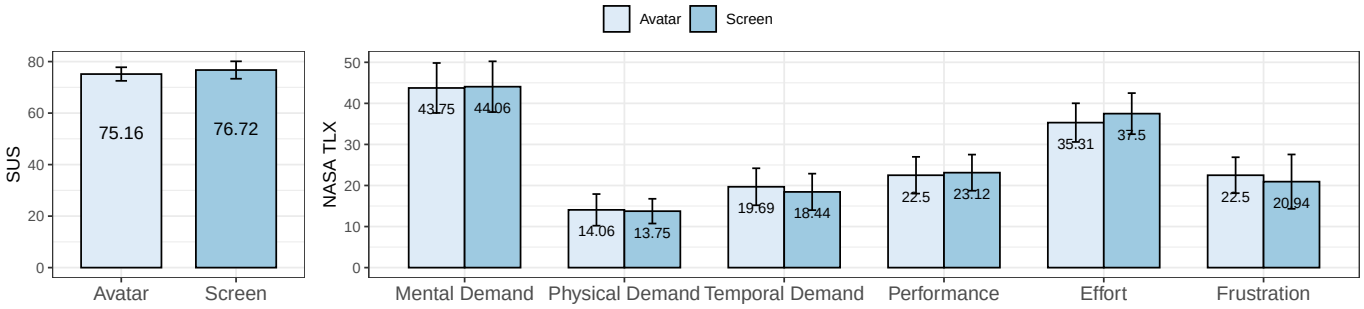

Fig. 5: Results of the questionnaires: (left) system usability scale (SUS) and (right) NASA task load index (TLX).

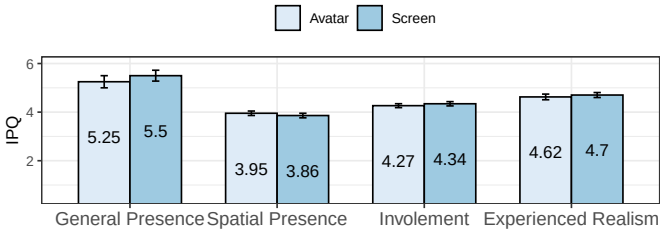

Fig. 6: Sense of presence results using *Igroup* presence questionnaire (IPQ) in the virtual environment.

1) Usability: The results of the usability assessment using the SUS questionnaire yielded an average score for *avatar* (M = 75.16, SD = 10.55) and *screen* (M = 76.72, SD = 13.53) (see Fig. 5). No significant differences were observed between the main effects and their interaction effect. Both configurations scored higher than 68, indicating above-average usability [52] and highlighting their potential usability benefits.

2) Task Load: The subjective task load was evaluated using an unweighted (raw) NASA-TLX questionnaire. Descriptive results, as shown in Fig. 5, demonstrated no significant differences in task load between configurations for all load dimensions (i.e., mental demand, physical demand, temporal demand, performance, effort, and frustration). Notably, mental demand (*avatar*: M = 43.75, SD = 24.32; *screen*: M = 44.06, SD = 24.71) emerged as the highest-rated load dimension, closely followed by effort (*avatar*: M = 35.31, SD = 18.83; *screen*: M = 37.50, SD = 20.00). This implies the proposed environment mainly impacts mental demand as participants engage with the virtual assistant to solve questions.

3) Presence: The sense of presence in the immersive VR environment was assessed using an IPQ questionnaire. Notably, no significant differences were observed among the configurations. Descriptive results indicated average scores for general presence (*avatar*: M = 5.25, SD = 1.00; *screen*: M = 5.50, SD = 0.89), spatial presence (*avatar*: M = 3.95, SD = 0.36; *screen*: M = 3.86, SD = 0.37), involvement (*avatar*: M = 4.27, SD = 0.32; *screen*: M = 4.34, SD = 0.35), and experienced realism (*avatar*: M = 4.62, SD = 0.46; *screen*: M = 4.70, SD = 0.41). Descriptively, general presence and experienced realism garnered the highest scores, followed by involvement and spatial presence (see Fig. 6).

C. Qualitative Participant Feedback (RQ3)

Participants offered open-ended feedback expressing their experiences about the difference between the two configurations of the study. Two participants preferred the *avatar* for heightened immersion. But there were suggestions regarding the improvement of voice synchronization in *avatar* condition. Participants also noted the need to pre-plan their questions, as delayed inquiries sometimes led to conversation interruptions by the system. Additionally, some participants occasionally forgot to release the talk button, causing the virtual assistant's on-screen response to be overridden by new text reacting to their continued speech.

V. Discussion

The proposed VR environment was evaluated in a pilot study to assess the feasibility and usability of a generative AI-based question-answering prototype with embodied and non-embodied virtual assistants.

RQ1 focused on the impact of the two virtual assistant configurations on user performance. We found that participants scored significantly higher when answering *knowledge-based* as compared to *analysis-based* questions in the *avatar* configuration. The fact that we did not find significant differences on *screen* leads us to believe that participants might have had an easier time keeping track of the virtual assistant's answers for the *analysis-based* questions in the *screen* than the *avatar* configuration.

Trends in descriptive results of the number of interactions, the tasks in *avatar* configuration require more interactions with the virtual assistant. This is supported by the fact that the participants had more interactions with the virtual assistant in the *avatar* than the *screen* with the *analysis-based* questions. Additionally, regarding user feedback, the participants did not show a clear preference for either the *screen* or the *avatar*. It could also indicate that they found benefits in both scenarios for different types of questions. As such, we can gain additional insights by gathering more information on the benefits of the *screen* compared to the *avatar* configuration, specifically considering the levels of cognitive complexity.

Knowing what makes each configuration helpful will allow us to combine the two configurations in a way that will be best for user performance in anatomy education.

RQ2 pertained to the impact of subjective measures on the virtual assistant configuration. There was no clear significance regarding usability, task load, or presence. Generally, it seemed that the participants were experiencing higher mental demand than physical demand. Additionally, the participants appeared to score higher for general presence than spatial presence. These preferences could be influenced by various factors, including the study location, experience with virtual assistants and VR, and potential distractions that prevented participants from forgetting their surroundings. Regarding usability, both configurations could qualify as relatively easy to use, which could add to our belief that each configuration has clear potential benefits. Therefore, combining both configurations could provide a full option for interacting with the environment.

RQ3 concerned the limitations and potential research directions of using generative AI for anatomy education. Integrating generative AI virtual assistant could adapt to users and provide personalized support [11]. It has the potential to offer an engaging, immersive learning experience, enhancing motivation. The generative AI-based virtual assistant can query a vast database of information and provide comprehensive information and resources according to the student's needs. However, evaluating the quality and accuracy of responses remains a crucial aspect that requires further investigation.

The results of this experiment indicate that generative AI appears more effective at providing direct answers, and the participants achieved higher scores on knowledge-based questions when cognitive complexity was low. However, analysis problems require more cognitive engagement, interpretation, and complex comprehension and may not be solved easily by the generative AI. Such questions involve human-like reasoning, complex judgment, and sometimes subjective interpretation. This shows that cognitive complexity, as measured by Bloom's taxonomy, is correlated with generative AI's ability to support learners. Generative AI may not solve analysis problems easily that demand human-like thinking, complicated judgment, and subjective interpretation, which require higher cognitive involvement, interpretation, and complex comprehension.

Limitations and Future Work: Our pilot study was conducted with 16 participants. Although enough for this study design to show sufficient power, it is still relatively small. During the study, we noticed that the way a participant phrased a question could impact the response they received. It correlates with other studies showing that virtual assistants could be biased, particularly when asked to find a relationship between items [56]. For AI-generated contents, there is a concern regarding generating somewhat inaccurate or misleading responses, the lack of transparency and unclear information on the data source used, and other related consequences, as noted in previous work [11]. Hence, future work should investigate the conversations' transcripts and the responses' accuracy. Regarding the verbal conversation using speech-to-text, some of our participants had problems being understood. A problem arose for participants with an accent and when they spoke too quickly. Consequently, these factors could have affected the participants' correctness scores and experience. The varying levels of experience for users using VR or virtual assistant, such as *ChatGPT* was found to be a challenge as well. Participants with little experience were less likely to interact with the environment and look around, which impacted how different they felt the two configurations were. For future studies, levels of experience should be considered as a covariate, or a balance of the groups based on their level of prior VR/virtual assistant experiences would be beneficial.

For future work, experience with a larger sample size and further development and integration with visual/object input and output, e.g., ChatGPT-4, could provide an extensive learning environment. Furthermore, providing tools to monitor learning progress and assessments may prove advantageous. Incorporating new modalities [57] and advanced techniques such as visual/acoustic emotion recognition [58]–[60], gaze engagement tracking [61], and body gesture analysis [62], [63] could improve the representation of virtual assistants. Additionally, it would be interesting to study the learning effects in a collaborative VR environment [64], [65].

VI. Conclusion

In this paper, we have developed an immersive VR environment featuring a generative AI-based embodied virtual assistant designed for human anatomy education. This environment was evaluated in a pilot user study involving 16 participants with no prior knowledge of the subject. The evaluation results demonstrated the impact of virtual assistant configurations on user performance. While there were small differences between the *avatar* and *screen*-based configurations in terms of the number of interactions, a significant difference emerged in the cognitive complexity level of questions associated with the *avatar*-based configuration. Additionally, we reported subjective measure results from usability, task load, and sense of presence. The combination of both virtual assistant configurations has the potential to offer a comprehensive solution for assisting and enhancing the learning experience. Moreover, our findings provide insights into potential benefits, limitations, and future research directions concerning the utilization of embodied virtual agents and generative AI conversational chatbots in education.

Acknowledgment

We thank Ishwar Kumar M.A. Sekaran for his contributions to study recruitment. We are also thankful to our study subjects for their support and participation. We wish to acknowledge the support from our sponsors, the National Science Foundation (2222661-2222663 and 2321274) and the National Institute of General Medical Sciences of the National Institutes of Health (P20 GM103446-E). This work was also prepared by LLNL under Contract DE-AC52-07NA27344. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the sponsors.

REFERENCES

[1] B. Preim and P. Saalfeld, "A survey of virtual human anatomy education systems," *Computers & Graphics*, vol. 71, pp. 132–153, 2018.

[2] C. F. Smith and B. McManus, "The integrated anatomy practical paper: A robust assessment method for anatomy education today," *Anatomical Sciences Education*, vol. 8, no. 1, pp. 63–73, 2015.

[3] L. W. Anderson and D. R. Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

[4] D. A. Morton and J. M. Colbert-Getz, "Measuring the impact of the flipped anatomy classroom: The importance of categorizing an assessment by bloom's taxonomy," *Anatomical sciences education*, vol. 10, no. 2, pp. 170–175, 2017.

[5] L. Chen, T. W. Day, W. Tang, and N. W. John, "Recent developments and future challenges in medical mixed reality," in *Proc. of IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 2017, pp. 123–135.

[6] V. Chheang, P. Saalfeld, F. Joeres, C. Boedecker, T. Huber, F. Huettl, H. Lang, B. Preim, and C. Hansen, "A collaborative virtual reality environment for liver surgery planning," *Computers & Graphics*, vol. 99, pp. 234–246, 2021.

[7] B. Preim, P. Saalfeld, and C. Hansen, "Virtual and augmented reality for educational anatomy," in *Digital Anatomy: Applications of Virtual, Mixed and Augmented Reality*. Springer, 2021, pp. 299–324.

[8] J. W. V. De Faria, M. J. Teixeira, L. d. M. S. Júnior, J. P. Otoch, and E. G. Figueiredo, "Virtual and stereoscopic anatomy: when virtual reality meets medical education," *Journal of neurosurgery*, vol. 125, no. 5, pp. 1105–1111, 2016.

[9] V. Chheang, V. Fischer, H. Buggenhagen, T. Huber, F. Huettl, W. Kneist, B. Preim, P. Saalfeld, and C. Hansen, "Toward interprofessional team training for surgeons and anesthesiologists using virtual reality," *International journal of computer assisted radiology and surgery*, vol. 15, pp. 2109–2118, 2020.

[10] N. Barari, M. RezaeiZadeh, A. Khorasani, and F. Alami, "Designing and validating educational standards for e-teaching in virtual learning environments (vles), based on revised bloom's taxonomy," *Interactive Learning Environments*, vol. 30, no. 9, pp. 1640–1652, 2022.

[11] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, 2023.

[12] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences*, vol. 13, no. 4, p. 410, 2023.

[13] Y. Chen, S. Jensen, L. J. Albert, S. Gupta, and T. Lee, "Artificial intelligence (ai) student assistants in the classroom: Designing chatbots to support student success," *Information Systems Frontiers*, vol. 25, no. 1, pp. 161–182, 2023.

[14] T. N. Fitria, N. E. Simbolon, and A. Afdaleni, "Chatbots as online chat conversation in the education sector," *International Journal of Computer and Information System (IJCIS)*, vol. 4, no. 3, pp. 93–104, 2023.

[15] M.-H. Hsu, T.-M. Chan, and C.-S. Yu, "Termbot: A chatbot-based crossword game for gamified medical terminology learning," *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, p. 4185, 2023.

[16] W. Tam, T. Huynh, A. Tang, S. Luong, Y. Khatri, and W. Zhou, "Nursing education in the age of artificial intelligence powered chatbots (ai-chatbots): Are we ready yet?" *Nurse Education Today*, vol. 129, p. 105917, 2023.

[17] G. Caldarini, S. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots," *Information*, vol. 13, no. 1, p. 41, 2022.

[18] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[21] S. B. Patel and K. Lam, "Chatgpt: the future of discharge summaries?" *The Lancet Digital Health*, vol. 5, no. 3, pp. e107–e108, 2023.

[22] A. Howard, W. Hope, and A. Gerada, "Chatgpt and antimicrobial advice: the end of the consulting infection doctor?" *The Lancet Infectious Diseases*, vol. 23, no. 4, pp. 405–406, 2023.

[23] K. Cheng, Q. Guo, Y. He, Y. Lu, S. Gu, and H. Wu, "Exploring the potential of gpt-4 in biomedical engineering: the dawn of a new era," *Annals of Biomedical Engineering*, pp. 1–9, 2023.

[24] Y. He, H. Tang, D. Wang, S. Gu, G. Ni, and H. Wu, "Will chatgpt/gpt-4 be a lighthouse to guide spinal surgeons?" *Annals of Biomedical Engineering*, pp. 1–4, 2023.

[25] S. Pedram, G. Kennedy, and S. Sanzone, "Toward the validation of vr-hmds for medical education: a systematic literature review," *Virtual Reality*, pp. 1–26, 2023.

[26] H. Mäkinen, E. Haavisto, S. Havola, and J.-M. Koivisto, "User experiences of virtual reality technologies for healthcare in learning: An integrative review," *Behaviour & Information Technology*, vol. 41, no. 1, pp. 1–17, 2022.

[27] J. J. Reyes-Cabrera, J. M. Santana-Núñez, A. Trujillo-Pino, M. Maynar, and M. A. Rodriguez-Florido, "Learning Anatomy through Shared Virtual Reality," in *Eurographics Workshop on Visual Computing for Biology and Medicine*. The Eurographics Association, 2022.

[28] R. Barmaki, K. Yu, R. Pearlman, R. Shingles, F. Bork, G. M. Osgood, and N. Navab, "Enhancement of anatomical education using augmented reality: An empirical study of body painting," *Anatomical sciences education*, vol. 12, no. 6, pp. 599–609, 2019.

[29] F. Bork, R. Barmaki, U. Eck, P. Fallavolita, B. Fuerst, and N. Navab, "Exploring non-reversing magic mirrors for screen-based augmented reality systems," in *2017 IEEE virtual reality (VR)*. IEEE, 2017, pp. 373–374.

[30] E. Kurt, S. E. Yurdakul, and A. Ataç, "An overview of the technologies used for anatomy education in terms of medical history," *Procedia - Social and Behavioral Sciences*, vol. 103, pp. 109–115, 2013, international Educational Technology Conference.

[31] C. Erolin, L. Reid, and S. McDougall, "Using virtual reality to complement and enhance anatomy education," *Journal of visual communication in medicine*, vol. 42, no. 3, pp. 93–101, 2019.

[32] K. Nakai, S. Terada, A. Takahara, D. Hage, R. S. Tubbs, and J. Iwanaga, "Anatomy education for medical students in a virtual reality workspace: A pilot study," *Clinical Anatomy*, vol. 35, no. 1, pp. 40–44, 2022.

[33] R. Kurul, M. N. Ögün, A. Neriman Narin, Ş. Avci, and B. Yazgan, "An alternative method for anatomy training: Immersive virtual reality," *Anatomical Sciences Education*, vol. 13, no. 5, pp. 648–656, 2020.

[34] J. Falah, S. Khan, T. Alfalah, S. F. M. Alfalah, W. Chan, D. K. Harrison, and V. Charissis, "Virtual reality medical training system for anatomy education," in *Science and Information Conference*, 2014, pp. 752–758.

[35] S. G. Izard and J. A. J. Méndez, "Virtual reality medical training system," in *Proceedings of the fourth international conference on technological ecosystems for enhancing multiculturality*, 2016, pp. 479–485.

[36] P. Saalfeld, A. Schmeier, W. D'Hanis, H.-J. Rothkötter, and B. Preim, "Student and Teacher Meet in a Shared Virtual Reality: A one-on-one Tutoring System for Anatomy Education," in *Eurographics Workshop on Visual Computing for Biology and Medicine (EG VCBM)*, 2020, pp. 55–59.

[37] D. Schott, P. Saalfeld, G. Schmidt, F. Joeres, C. Boedecker, F. Huettl, H. Lang, T. Huber, B. Preim, and C. Hansen, "A vr/ar environment for multi-user liver anatomy education," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2021, pp. 296–305.

[38] C. Moro, Z. Štromberga, A. Raikos, and A. Stirling, "The effectiveness of virtual and augmented reality in health sciences and medical anatomy," *Anatomical sciences education*, vol. 10, no. 6, pp. 549–559, 2017.

[39] R. Q. Mao, L. Lan, J. Kay, R. Lohre, O. R. Ayeni, D. P. Goel *et al.*, "Immersive virtual reality for surgical training: a systematic review," *Journal of Surgical Research*, vol. 268, pp. 40–58, 2021.

[40] A. Bernardo, "Virtual reality and simulation in neurosurgical training," *World neurosurgery*, vol. 106, pp. 1015–1029, 2017.

[41] N. Norouzi, K. Kim, J. Hochreiter, M. Lee, S. Daher, G. Bruder, and G. Welch, "A systematic survey of 15 years of user studies published in the intelligent virtual agents conference," in *Proceedings of the 18th international conference on intelligent virtual agents*, 2018, pp. 17–22.

[42] R. Barmaki and C. E. Hughes, "Embodiment analytics of practicing teachers in a virtual immersive environment," *Journal of Computer Assisted Learning*, vol. 34, no. 4, pp. 387–396, 2018.

[43] K. Yu, R. Barmaki, M. Unberath, A. Mears, J. Brey, T. H. Chung, and N. Navab, "On the accuracy of low-cost motion capture systems for range of motion measurements," in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, vol. 10579. SPIE, 2018, pp. 90–95.

[44] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch, "Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2018, pp. 105–114.

[45] S. Haesler, K. Kim, G. Bruder, and G. Welch, "Seeing is believing: Improving the perceived trust in visually embodied alexa in augmented reality," in *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2018, pp. 204–205.

[46] K. Kim, C. M. de Melo, N. Norouzi, G. Bruder, and G. F. Welch, "Reducing task load with an embodied intelligent virtual assistant for improved performance in collaborative decision making," in *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2020, pp. 529–538.

[47] H. Yao, A. G. de Siqueira, A. Bafna, D. Peterkin, J. Richards, M. L. Rogers, A. Foster, I. Galynker, and B. Lok, "A virtual human interaction using scaffolded ping-pong feedback for healthcare learners to practice empathy skills," in *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, 2022, pp. 1–8.

[48] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior research methods*, vol. 39, no. 2, pp. 175–191, 2007.

[49] H. Kenngott, J. Wünscher, M. Wagner, A. Preukschas, A. Wekerle, P. Neher, S. Suwelack, S. Speidel, F. Nickel, D. Oladokun *et al.*, "Openhelp (heidelberg laparoscopy phantom): development of an open-source surgical evaluation and training tool," *Surgical endoscopy*, vol. 29, no. 11, pp. 3338–3347, 2015.

[50] M. Feick, N. Kleer, A. Tang, and A. Krüger, "The virtual reality questionnaire toolkit," in *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 68–69.

[51] J. Brooke, "SUS: A quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, 1995.

[52] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *Journal of usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

[53] S. G. Hart, "NASA-Task Load Index (NASA-TLX); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, 2006, pp. 904–908.

[54] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence: Teleoperators & Virtual Environments*, vol. 10, no. 3, pp. 266–281, 2001.

[55] V. Schwind, P. Knierim, N. Haas, and N. Henze, "Using presence questionnaires in virtual reality," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–12.

[56] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in chatgpt: implications in scientific writing," *Cureus*, vol. 15, no. 2, 2023.

[57] J. Li, V. Chheang, P. Kullu, E. Brignac, Z. Guo, A. Bhat, K. E. Barner, and R. L. Barmaki, "Mmasd: A multimodal dataset for autism intervention analysis," pp. 397–405, 2023.

[58] A. Hartholt, E. Fast, A. Reilly, W. Whitcup, M. Liewer, and S. Mozgai, "Ubiquitous virtual humans: A multi-platform framework for embodied ai agents in xr," in *IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019, pp. 308–3084.

[59] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593–617, 2022.

[60] J. Li, A. Bhat, and R. Barmaki, "A two-stage multi-modal affect analysis framework for children with autism spectrum disorder," in *Proceedings of the AAAI-21 Workshop on Affective Content Analysis*, 2021, pp. 1–8.

[61] Z. Guo, V. Chheang, J. Li, K. E. Barner, A. Bhat, and R. Barmaki, "Social visual behavior analytics for autism therapy of children based on automated mutual gaze detection," in *Proceedings of the International Conference on Cooperative and Human Aspects of Software Engineering*, ser. CHASE '23, 2023.

[62] R. Barmaki and C. Hughes, "Gesturing and embodiment in teaching: Investigating the nonverbal behavior of teachers in a virtual rehearsal environment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[63] J. Li, A. Bhat, and R. Barmaki, "Pose uncertainty aware movement synchrony estimation via spatial-temporal graph transformer," in *Proceedings of the International Conference on Multimodal Interaction*, ser. ICMI '22, 2022, p. 73–82.

[64] A. Scavarelli, A. Arya, and R. J. Teather, "Virtual reality and augmented reality in social learning spaces: a literature review," *Virtual Reality*, vol. 25, pp. 257–277, 2021.

[65] V. Chheang, D. Schott, P. Saalfeld, L. Vradelis, T. Huber, F. Huettl, H. Lang, B. Preim, and C. Hansen, "Towards virtual teaching hospitals for advanced surgical training," in *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2022, pp. 410–414.