2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3503-2838-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISMAR59233.2023.00110

Supporting Co-Presence in Populated Virtual Environments by Actor Takeover of Animated Characters

Jingyi Zhang*
University College London

Klara Brandstätter †
University College London

Anthony Steed‡
University College London

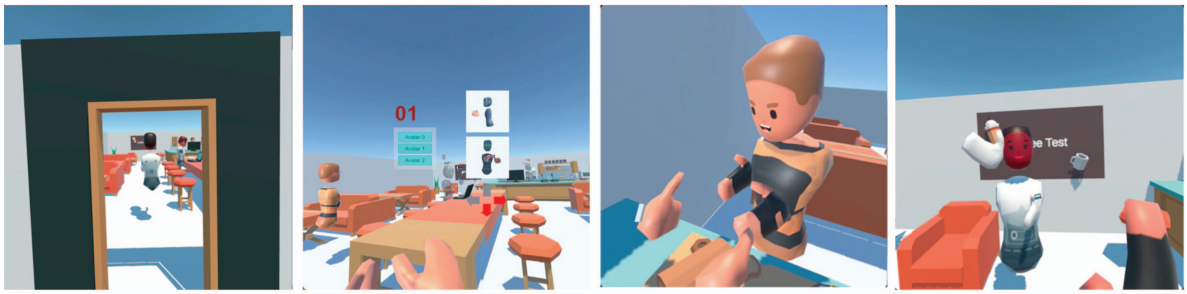

Figure 1: Scenario with one actor jumping between several "agent avatars" and interacting with the user when necessary. (a) The user enters the scene and discovers three agent avatars. All three agent avatars are playing pre-recorded, loopable clips. Two of them portray customers while the third one acts as the barista. An actor, who is invisible to the user, is ready to take over any of the agent avatars and interact with the user. (b) The actor realizes that the participant intended to interact with agent avatar number 0, who is playing the role of a barista. Consequently, the actor plans to take control of that agent avatar and embody it to respond to and interact with the user. (c) The actor takes control of Avatar0 and is now interacting with the user. Meanwhile, all the other agent avatars continue to play their pre-recorded clips. (d) The actor releases control of Avatar0 and switches to another agent avatar to interact with the user. This creates the illusion that all three agent avatars were being controlled by real humans.

Abstract

Online social virtual worlds are now becoming widely available on consumer devices including virtual reality headsets. One goal of a virtual world could be to give a user an experience of a crowded environment with many virtual humans. However, gathering enough personnel to control the necessary number of avatars for creating a realistic scene is usually difficult. Additionally, current technology is not capable of fully simulating avatars with behaviours, especially when interaction with users is required. In this paper, we develop a system that enables an actor to take over control of one of a set of avatars. We built an immersive interface that allows an actor to select an avatar to take over and then segue into the currently playing animation. By allowing one person to take control of multiple avatars, we can enhance the plausibility of environments inhabited by simulated characters. In an experiment, we show that in a cafe scenario, one actor can take over the roles of a barista and two customers. Experiment participants reported experiencing the scene as if it were populated by more than one actor. This system and experiment demonstrate the feasibility of one actor controlling multiple avatars sequentially, thus enhancing users' feelings of being in a populated environment.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

*e-mail: jy.zhang@ucl.ac.uk
†e-mail: k.brandstatter@ucl.ac.uk
‡e-mail: a.steed@ucl.ac.uk

1 Introduction

The advancement of consumer virtual reality (VR) systems and high-speed internet services has enabled a broad range of new social virtual reality (SVR) applications. These range from expansive social applications such as VRChat[1] to more curated and managed participatory experiences such as those from Tender Claws[2]. Current platforms tend to support relatively small numbers of users in a room. For example, MeetinVR[3] supports up to 32 simultaneous users and VRChat[1] supports up to 16 players in the same hub.

To create experiences that appear to be densely populated we can augment users with avatars representing autonomous agent simulations. Such agents might have a similar appearance as user avatars and thus the user might expect them to have at least some reactive and communicative behaviours. Since the plausibility of the characters' behaviour can significantly affect the level of immersion and presence experienced by the user [34], building autonomous intelligent agents has been an area of interest for decades. Trials have been made to simulate human behaviour by using generative models [27] or reinforcement learning and procedural animation [7], which give the agents the ability to use language [27], have emotion [7] and act according to aims and goals [7, 27].

Despite these advances, purely automatic scene population still falls short of generating believable open-ended scenarios for a user to enter. The enormous freedom that VR affords users to interact, gesture and speak has in turn posed huge challenges in creating believable intelligent characters. The difficulty of this task is at least as hard as generating fully believable language, not to mention the

[1]VRChat: https://hello.vrchat.com/
[2]Tender Claws: https://tenderclaws.com/
[3]MeetinVR: https://www.meetinvr.com/

979-8-3503-2838-7/23/$31.00 ©2023 IEEE
DOI 10.1109/ISMAR59233.2023.00110

uncanny valley effect that even short animation can solicit [29] and the variability and intricacies of non-verbal behaviours. Passing a VR version of the Turing test [33] is still a long way off.

Inspired by the idea from Neal Stephenson's book "The Diamond Age" [31] where characters are played by hired "ractors", we expect that an actor might be able take over an agent to fill in the gaps in its reactive behaviours. Thus, as an intermediate solution to enhance the immersion of users in social VR, we developed a takeover system in which agents could be controlled by a human when needed whilst pre-recorded motions were played autonomously at other times. In the scenario where only one participant was present in the scene, we could allow a single actor to jump amongst multiple agents. Since agents are virtual humans driven by computer algorithm whereas avatars are controlled by humans [2], we named the agents occasionally controlled by the actor - "agent avatars".

The takeover system was built on the Ubiq social VR platform [13]. We created a cafe scenario populated with three agent avatars. Both actor and participant could enter this cafe using Meta Quest 2[4]. The actor was able to see all agent avatars available for takeover. Each agent avatar displayed with a future keyframe selected from the pre-recorded loopable clips. The actor was asked to mimic the motion cued by the keyframe of the agent avatar. A linear blending would produce animation from the keyframe avatar to the actor's current position and posture to compensate for the difference when takeover happened. The actor's avatar was normally invisible to the participant and could only be "seen" when one of the existing agent avatars had been taken over. In this way, the participant was unaware of the transfer of control and an illusion that all characters in the scene have human intelligence was produced.

In an experiment, we asked participants to perform a sequence of tasks in a cafe scenario while the actor took over different agent avatars and interact with participants. We investigated the social presence using a questionnaire from Harms and Biocca [17] and conducted interviews to gather more in-depth feedback on the experiences as well as suggestions for improvement. The results proved that our system improved the perceived plausibility of the virtual environments inhabited by pre-recorded agent avatars.

Apart from its entertainment use-cases in social VR games or immersive film experiences, the takeover system holds vast potential for applications in training scenarios. Enhanced by within-world intervention from an instructor who can seamless take control of any character simulated in scene, the system could be further developed with simulated virtual students (E.g TeachLivE^{TM}) and patients for teacher and doctor training. Moreover, the system could be a valuable tool for psychological experiments involving crowd and avatar interactions, eliminating the need of hiring numerous actors while maintaining variable control.

In summary, the aim of our work is to enhance environments populated by non-interactive characters by letting an actor take control of several agent avatars whenever interaction with participant is necessary. Specifically, this paper provides the following contributions:

- A system that allows a single actor to jump between multiple agent avatars and inhabit them in a virtual reality environment.
- A subsequent experiment to prove the utility of the system in creating the impression of a more populated environment.
- Some initial findings and guidelines in how an actor can successfully take over several characters.

2 Related Work

Virtual reality technology enables a new way of face-to-face conversation, allowing close to real-life social interaction to take place regardless of distance. Social presence or co-presence, which refers to the experience of "being together" with another social being [3], holds significant importance in VR studies involving social interaction due to the fact that social reactions become more realistic and close to real-life human-human interactions when users have a greater sense of social presence [19]. To assess social presence researchers have employed subjective self-report measurements and behavioural indicators [3]. In VR experiments, the former is more commonly used and has been well-developed in the form of questionnaires [17, 37].

Social interactions not only take place between user controlled avatars but may also happen between users and agents. Creating believable virtual characters has always been an appealing task due to its wide usage in various fields from entertainment to psychology experiments. The term believability describes how a character's behaviour matches the expectations of the viewer [7]. Maintaining the believability of a virtual character is one of the crucial factors in creating a plausible interaction with it. The plausibility illusion, proposed by Slater [28], describes the illusion that the event in VR is really occurring. When it comes to interaction with virtual characters, achieving the plausibility illusion needs to fulfil three criteria: the characters should demonstrate responsiveness to the participant's behaviour: their behaviour should refer directly to the participant; and their behaviour should align with what the participant will expect in similar real-world scenarios [14].

Large language models, such as ChatGPT, which generate responses closely resembling natural human language have recently emerged in quick succession with remarkable performance [23]. However, when it comes to generating non-verbal behaviours that express body language, there is still room for improvement. A virtual character's non-verbal behaviour can be modelled using statistical models, rule-based methods or machine learning [9]. Statistical models apply the probability distributions generated from real-life human-human interaction data to the virtual character's behaviour [21, 25]. This method, however, can only generate behaviour in a specific domain. It can not build up to a versatile virtual character that can fit into complex scenarios and perform natural social interactions. Rule-based methods utilize rules derived from real-life human interaction, allowing pre-captured or existing motions to be played back accordingly. The playback can be triggered either by a pre-programmed algorithm [4, 24], or similar to a "Wizard-of-Oz" system [26], by the experimenter using a set of buttons. However, this approach has limitations in terms of motion variety, which can lead to repetition and lack of naturalness over time. Recent research primarily focused on machine learning algorithms to model non-verbal behaviours. These models were usually trained from large data sets. They not only had the ability to animate facial expression [11] and gesture [12], but also can simulate different behaviours influenced by factors including emotions [7, 15], goals [7], or user's behaviour as the conversation flows [8].

Having a single character capable of interacting with users is not sufficient to create the illusion of being in a complete virtual world. As in real life, the world might be composed of many individuals. Studies of crowd simulation mostly focused on the overall behavioural realism of the crowd, such as movement [35]. In that study, each individual only performed simple behaviours such as walking, as opposed to the diverse range of behaviours we have in the real world. Collecting data to study crowd behaviours is challenging due to numerous reasons such as ethical concerns and technical limitations. To overcome this problem and boost the dataset available, Yin et. al [36] proposed the one-man-crowd paradigm, where a single actor repeatedly added new motions until the entire crowd was recorded. This method thus suggests the potential to gather full-body motions from an individual actor to form multiple agents that populate the scene. In contract to that work, actors in our scenario must take over existing animations that are already playing.

In order to maintain a believable scenario, we can see a need for

[4]Meta Quest 2: https://www.meta.com/quest/products/quest-2/

an actor to be able to intervene. For example, to give an answer to a question the participant makes that the designers had not anticipated. Further, we can also see that while it may be possible in the future to create very believable characters with rich back stories and varied interaction capabilities, there will still be a need for a director or actor to intervene to create novel situations or interventions to create a more plausible experience, in a similar manner to a games master in role-playing games or director in a live television studio. In particular, in a social VR scenario an actor might want to take over the behaviour of an agent avatar that is already in the scene rather than have a new agent avatar join the scene.

An example demonstrate the above need could be a current successful commercial system, TeachLivE^{TM} [5]. TeachLivE^{TM} [10] is a teacher training system which allows one interactor to puppeteer multiple virtual students. Differing from this system, we aim to provide full control of avatar movement, allowing seamless take over from an ongoing animation through matching the movement.

In order to let an actor take control over an agent avatar, we need to inform them about the avatar's future position and intended pose. VR systems designed to lead users to imitate others' postures have been widely researched in sports coaching. In these systems, the feedback to learners can be divided into eventual ones and immediate ones. The former can be a score report that concludes the overall performance of the learner [5, 22]. The latter, instant feedback, can be categorised into changes in coloured joint points [22] or skeleton [5, 18, 38], virtual coaches placed around the learner's avatar [6, 22] and translucent or coloured virtual coach avatar overlapping the learner's one [18, 22, 38]. Mirrors [1] and third-person viewpoints [16] have also been used to assist the learner to adjust their posture. To highlight the important postures that need to be matched, the keyframes from the pre-recorded movement of the virtual coach could be selected and presented to the learner [1]. The key difference to most of these previous works, is that in our system we cannot pause the animation to allow the actor to reach the correct position.

3 System Design

This section introduces the takeover system which enabled the actor to take control of an agent avatar when necessary. This is a proof-of-concept prototype to demonstrate that the takeover process is achievable for an actor with some training. In Section 4, we show that the resulting system works by demonstrating how participants react to live scenarios with an actor. However, the takeover system itself is novel and no doubt can be improved in the future.

A scene of non-interactive agent avatars was created. These agent avatars played back pre-recorded animation clips of them performing various actions relevant to their roles. With this setup, the aim of the actor was as follows: interrupt the pre-recorded loop when a participant attempted to interact with one of the agent avatars; take over control and interact with the participant; and then release control to let the agent avatar resume its previous recorded actions. The entire process should run smoothly and continuously, without the participant detecting any change in control or unnatural movement. In particular the agent avatar must not appear to freeze.

To ensure a seamless takeover process, the actor needed to expeditiously choose a suitable target avatar and match its posture and position as closely as possible before starting to control. The actor should be aware of the available options of agent avatars, the avatars' positions and poses, and the position and facing direction of the participant's avatar. The system should enable the actor to select an agent avatar, teleport to its position, adjust their own pose to match that of the agent avatar's one, and take over and release as required.

The system was built on the Ubiq framework [13], an open-source Unity networking library, which facilitates object spawning, message passing, avatar management and lightweight XR interactions including grasping and releasing objects.

[5]https://sites.google.com/view/teachlive

Below we describe a system designed to assist the actor in acquiring the necessary information for the takeover process and providing guidance and hints on matching their own position and posture to the agent avatar going to be taken over.

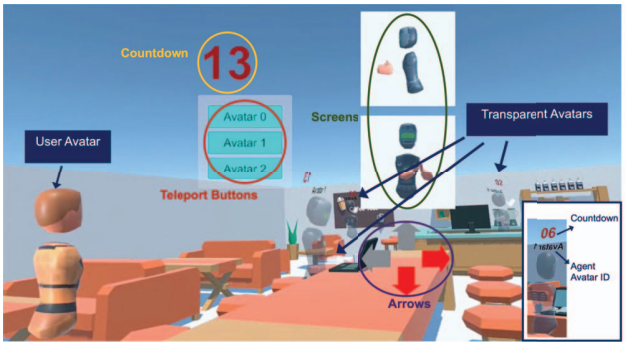

Figure 2: First-person view from the actor in VR. Transparent avatars: available choices of agent avatars to take over; In the centre: control panel (countdown, teleport buttons, two screens and four arrows) fixed in view, further explained in section 3.3; Right-bottom corner: A zoom in of a transparent avatar with agent avatar ID and countdown. The countdown indicated the time left before it jumped to the next future keyframe position, which is further explained in section 3.2

3.1 Avatars in Scene

When entering the scene, the actor could find several transparent avatars with IDs, which represented the options they were able to take over control, as shown in Fig. 2. The participant's avatar was visible to the actor and could assist in determining which agent avatar to take control of next.

3.2 Keyframe Extraction

The continuous motion played in the participant's view offers no opportunity for the actor to jump in and takeover. Therefore to allocate time for the actor to match the posture and position of the agent avatars we extracted keyframes from the motion recordings. The keyframes were represented as transparent avatars which would appear to be temporarily frozen in the actor's view only (i.e. not shown in the participant's view). A countdown placed on the top of the agent avatar's head indicated the remaining time before it jumped to the next future keyframe position. When the countdown reached zero, the agent avatar in the participant's view arrived at the temporarily frozen position and posture displayed as transparent avatar in the actor's view. With this setting, if the actor took over the agent avatar towards the end of the countdown, the actor's own avatar should have a very close posture and position to the one interrupted from the recording in the participants' view.

Keyframe selection was based on two factors. Firstly, the moving velocity and acceleration should be relatively low. When getting ready to take over control, the actor was required to stay static. If the takeover happened when the agent avatar was having a relatively high speed, it would result in an unnatural sudden stop. Secondly, the time interval between the current and the next keyframe should be within approximately 10 seconds. This was to ensure the actor has sufficient time to select the desired agent avatar, take control of it, and that the posture of the agent avatar was not too different from the temporarily frozen posture due to a long time difference.

3.3 Control Panel

The basic control panel was fixed in the actor's view, as shown in Fig. 2. It included multiple buttons, a countdown timer, two screens, and four arrows located in the left-bottom corner. Each button on the

panel corresponded to the avatar ID above each transparent avatar. Clicking on the button teleported the actor to the chosen transparent avatar's position and rotation. The countdown fixed in the actor's view matched that of the nearest transparent avatar, providing a reference for the actor when attempting to take over control. For example the actor could assess whether they had enough time to get into position. The two screens displayed real-time views from two cameras set in front of and on the left side of the actor's avatar. To provide the actor with a clearer view, the screens displayed no environmental models but only the actor's avatar and the transparent avatar, with the latter rendered in blue colour. The four arrows started as grey but turned red to indicate the head rotation direction required to match the transparent avatar.

3.4 Match Guidance and Hints

When part of the actor's avatar matched the transparent avatar within a predefined range, this body part of the transparent avatar turned green, as shown in Fig. 3. The actor was able to observe the colour change on both the transparent avatar and the one on the two screens fixed in their view.

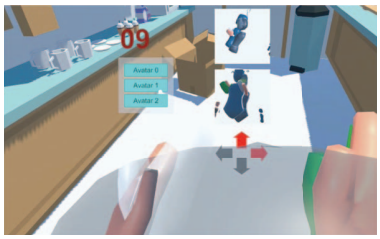

Figure 3: The actor was trying to match the agent avatar in order to take over its control. The colour change on the right hand of the avatar indicated the match was within an acceptable range.

3.5 Takeover and Release

Once all body parts of the agent avatar were matched, the actor could take over control with a single click on their controller. While the actor was in control, all other transparent avatars and the control panel were hidden from view. The match did not need to be perfect since the takeover might happen before the countdown ends so the agent avatar in the participant's view had not yet reached the position and posture indicated by the transparent avatar. Instead, a smooth linear interpolation method blended the current posture and position of the actor's avatar with the position where this agent avatar stopped playing the pre-recorded clip in the participant's view. The linear change of position and rotation was calculated and an animation of this blending process was played on the participant's side so that they would not recognise the transition. While the blending animation was playing, the actor's view turned black and they were asked to stay still and get ready to interact with the participant once their view came back.

To release control of an agent avatar, the same button on the controller needed to be clicked. The blending process worked the same but in the opposite direction. This time the position where the actor released the agent avatar was blended with the position where the pre-recorded clip was interrupted. The clips would continue to play in a loop from where it was interrupted in the participant's view. The actor's view turned black while the blending animation was playing. Once the blending was done, the actor's view was back to normal with all transparent avatars and the control panel reappeared.

The whole procedure of taking over is shown in Fig. 4. Firstly, the actor made the decision on which agent avatar to take over. This could be done with the assistance of observing the participant's avatar and predicting which agent avatar they are going to interact with. Once the decision was made, the actor clicked on the teleport button of the accorded agent avatar to be teleported to the transparent avatar's position and rotation. The actor then needed to match the transparent avatar's two hands and head before the countdown ended with the help of colour change, two screens and arrows. When all parts of the avatar were matched, a text would pop out, instructing the actor to press the button on the controller to take over control of the agent avatar. Once the takeover button on the controller was pushed, the blending animation started to play in the participant's view and the actor should get ready to interact with the participant after their view turned back from black.

4 Experiment

An experiment was conducted to investigate how participants perceive the periodically human-controlled avatars (avatars which were controlled by humans only when interaction happened) and whether our method enhanced the plausibility of the scene inhabited by non-interactable virtual characters. We made a comparison between the interaction produced by our system and the one produced by playing pre-recorded responses. The latter was using a "Wizard-of-Oz" setup

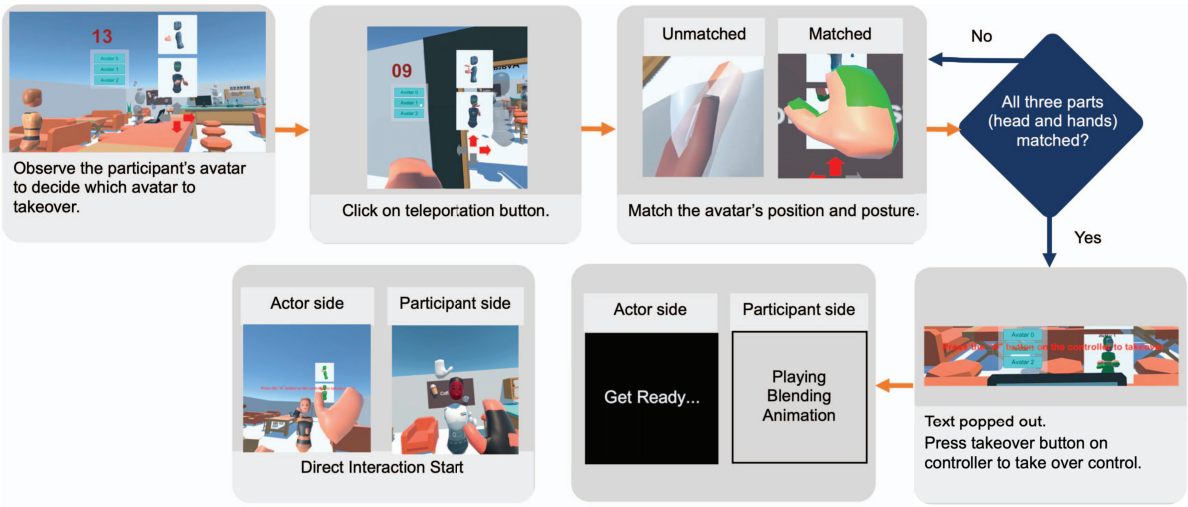

Figure 4: Takeover procedure

Figure 5: Layout of the Cafe and the Waiting Room

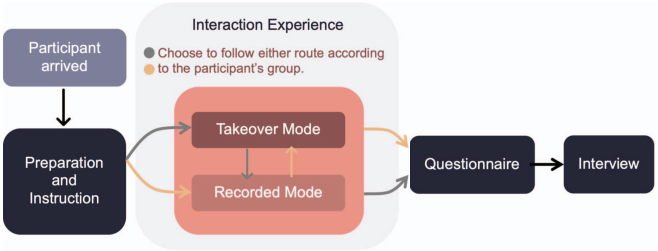

Figure 6: Procedure of the experiment

where the experimenter played the pre-recorded responses when the participants interacted with the avatar to imitate some form of algorithmic response system. Thus the two conditions were: interacting with the avatars truly taken over by the actor, which is hereafter referred to as "Takeover Mode"; or interacting with the avatars that had the pre-recorded responses, which is hereafter referred to as "Recorded Mode".

Based on the aim and goals of our system, we made the following hypotheses about the experiment:

- H1: Social presence of participants will be higher in Takeover mode versus Recorded mode.
- H2: The perceived number of real human controlled avatars will be higher in Takeover mode versus Recorded mode.

20 participants (7 males) ranging in age from 19 to 26 were invited for our experiment. They all had a normal or corrected-to-normal vision. The experiment was conducted in the Immersive Virtual Environments Laboratory in the Computer Science Department at University College London. The study was run under ethics approval by the UCL Research Ethics Committee (approval ID 4547/012).

4.1 Scenario design

We designed a scene with a waiting room and a cafe, as shown in Fig. 5. The cafe was smaller than the guardian that was set up, as a 6x8m space.

Three agent avatars, a barista and two customers, were placed in the scene for the participants to interact with. Two clips were pre-recorded. One would be played in a loop when not interacting with participants. The other would only be played by the experimenter when the user tried to interact with the agent avatar in the comparison group. The clips were made using a Meta Quest[4] and a record and replay tool in Ubiq [30]. A motion capture person was told about their mission and identity in the scene. They entered the scene and started performing according to their roles. In the meantime, the position and rotation of their head and hands were recorded.

The role of Avatar0 was a bartender. For the recording of the loop-able clip, the motion capture person was asked to enter the bar using the staff access, wash their hands, inventory the milk bottles and load the number onto the system using the checkout machine, and then exit the bar counter. For the recording of the response clip, the motion capture person was asked to lead the user to the dessert table and point to the doughnut. The motion capture person was told that this recording will be used as a response to the user asking them for a free doughnut.

The role of Avatar1 was a customer. For the recording of the loopable clip, the motion capture person was asked to work with the laptop and type using the keyboard. For the recording of the response clip, the motion capture person was asked to thank users by clapping their hands. The motion capture person was told that this recording will be used as a response to the user helping to pour milk into their cup.

The role of Avatar2 was also a customer. For the recording of the loopable clip, the motion capture person was asked to walk into the cafe through the customer entrance, walk up to the board with "Coffee Test" written on it, pretend to read the contents on the board carefully, and then leave the cafe. For the recording of the response clip, the motion capture person was asked to say "hi" to the user by waving their hands. The motion capture person was told that this recording will be used as a response to the user greeting them.

The participants were directed to interact with the agent avatars through assigned tasks. They were trained to teleport using their controller and could freely navigate the scene by walking in reality. Participants were provided with pre-set objects to interact with, including grabbing cups, milk bottles, and doughnuts, as well as pouring milk into the cups.

4.2 Conditions and tasks

As introduced, there were two conditions designed: Takeover Mode and Recorded Mode. Participants were divided into two groups, one experiencing the first condition followed by the second, and the other vice versa.

During the experiment, participants were asked to complete three tasks for both conditions in VR. The participants were informed that in the upcoming experiment, several avatars would inhabit the empty cafe they just saw. They were told that while they could communicate verbally with other avatars, the avatars would only respond to them using body language. The participants were given three tasks in order. First, say "hi" to the avatar in front of the coffee test board. Second, take the milk bottle from the bar, pour the milk into the cup beside the laptop, and then drop the bottle into the box in front of the bar. Third, ask the barista for a free doughnut, grab the doughnut and walk out of the cafe. The participants were not told the aim of the experiment.

4.3 Experiment procedure

The whole procedure of the experiment is illustrated in Fig. 6. The experiment consists of four parts: Preparation and instruction, Interaction Experience, Questionnaire, and Interview.

4.3.1 Preparation and Instruction

On arrival at the lab, the participants were asked to read through a participant information sheet and sign a consent form. They were then guided by the experimenter to put on the Meta Quest 2. They were asked to walk around in the default Welcome Lobby and adjust the HMD until it was sitting comfortably and the view was clear. After that, the experimenter instructed the participants to launch the app and wait in the waiting room and introduced them to the basic user interaction and moving methods with the Quest 2. The participants were then asked to enter the virtual cafe to practice the interactions they had just been taught. They were guided through the procedure of the most complex task (all three tasks will be further introduced later in this section), pouring milk, step-by-step. After they felt familiar with the environment, they were asked to quit the app and get ready for the formal experiment.

4.3.2 Interaction Experience

The participants were asked to launch the app and wait in the waiting room until they were told to enter the cafe and start doing the tasks. When the participant was in the waiting room, the experimenter would set the agent avatars to the corresponding mode (either Takeover mode or Recorded mode). The experimenter would either become an actor to take over the corresponding agent avatar in turn and respond to participants as they would do in real life; or play the pre-recorded responses for each agent avatar when the participant was trying to interact with it. Once the participants had finished all three tasks, they were asked to quit and re-enter the app and perform the same tasks again with another condition.

4.3.3 Questionnaire

After they finished the experience for both conditions, the participants were asked to fill in a questionnaire about social presence and basic information. The questionnaire consists of three parts: a social presence part repeated once for each condition and general information about the participants. The social presence part referenced the Networked Minds Measure of Social Presence Inventory proposed by Harms et al. [17], taking its co-presence and perceived behavioural interdependence sections for our study. The general information part asked participants to provide their gender, their prior experience with VR and the average time (in hours) they spent on video games per day in the last month. These were to investigate if there were any influential factors that contribute to the results.

4.3.4 Interview

Additionally, the participants were asked to complete an interview. During the interview, the participants were asked to discuss their feelings while interacting with the other avatars and their perception of being with another user like them who controlled an avatar in each condition. In the following, we show some sample questions that were asked during the interview:

- How many real-human controlled avatars do you think are in the scene for each condition? Real-human-controlled avatars mean the avatar is controlled by another user like you. It should be in the range of 0 to 3 for each turn.
- What makes you think the avatar is real-human controlled or not? What makes you feel suspicious? Please comment on each condition separately.
- Will you treat the 3 avatars as separate individuals? Please comment on each condition separately.
- Participants were told that there are 3 or 0 avatars controlled by real humans in each condition, then they were asked: Do you feel surprised about that?
- Participants were told that the 3 avatars were actually taken over in turn by one person only, and then they were asked: Do you feel surprised about that? Have you noticed the jump of the actor from one avatar to another? Have you noticed any unnatural movement when interacting with the avatar that made you feel the actor is taking over or releasing?

We provide a video sample showcasing the experiment procedure for both conditions in both actor's and participant's views in VR.

5 Results

5.1 Questionnaire

The questionnaire investigated the basic information about the participants and the social presence of the interaction in two sections: co-presence (CP) and perceived behavioural interdependence (PBI) (see [17]). Each section consisted of 6 questions which were scored on a scale of 1 to 5. The total score of each section was in the range of 5 to 30.

5.1.1 Takeover Mode vs. Recorded Mode

To test our hypothesis about social presence in two modes described in the experiment design section, a one-tailed paired t-test was conducted.

The CP score reported by the participants in Takeover mode ($M = 27.400, SD = 2.563$) was higher than the one in Recorded mode ($M = 24.750, SD = 4.689$). The same trend was observed for PBI score between Takeover mode ($M = 26.400, SD = 2.644$) and Recorded mode ($M = 23.250, SD = 4.363$). The takeover system in Takeover mode elicited a significant increase in the social presence scores, both for CP score ($M = 2.650$, 95%CI [0.267, 5.033], $t(19) = 2.328$, $p = 0.016$) and PBI score ($M = 3.150$, 95%CI [0.948, 5.352], $t(19) = 2.994$, $p = 0.004$), compared to the Recorded mode. Therefore, our hypothesis H1 about social presence can be accepted. Interacting with the Takeover mode enhanced the social presence perceived from purely recorded agent avatars in Recorded mode.

5.1.2 Influential factors

To better understand the result, we assessed two main influences: interaction order, and prior VR experience, which might influence participants' responses.

Interaction Order A two-way mixed ANOVA test was performed to investigate whether changes in social presence scores over different interaction modes varied for different interaction orders. Fig. 7 shows the score in CP and PBI section for each interaction mode in different interaction orders. Because the original CP score violated the assumption of the homogeneity of variances and covariances, it was transformed using reflect and square root to comply with the requirements for the ANOVA test [20].

There was a significant interaction between interaction order and interaction mode on both transformed CP score ($F(1,18) = 6.342$, $p = 0.021$, partial $\eta^2 = 0.261$) and PBI score ($F(1,18) = 5.708$, $p = 0.028$, partial $\eta^2 = 0.241$).

No significant differences were found in transformed CP score and PBI score between different interaction order in both interacting with Takeover mode (CP: $F(1,18) = 2.705$, $p = 0.117$, partial $\eta^2 = 0.131$; PBI: $F(1,18) = 0.444$, $p = 0.513$, partial $\eta^2 = 0.024$) or Recorded mode (CP: $F(1,18) = 2.252$, $p = 0.151$, partial $\eta^2 = 0.111$; PBI: $F(1,18) = 4.201$, $p = 0.055$, partial $\eta^2 = 0.189$). Thus, in the same interaction mode, different interaction order does not influence the result.

When interacting with Recorded mode first, the transformed CP score was significantly lower in Takeover mode compared to Recorded mode ($M = 1.052$, $SE = 0.416$, $p = 0.032$). Therefore, when interacting with Recorded mode before Takeover mode, the original CP score in Takeover mode was significantly higher than the one in Recorded mode. The same trend was found for PBI score as well. When interacting with Recorded mode first, the PBI score was significantly higher in Takeover mode compared to Recorded mode ($M = 5.400$, $SE = 1.790$, $p = 0.015$). However, the effect of mode was not significant when interacting with Takeover mode first (CP: $F(1,9) = 0.076.$, $p = 0.789$, partial $\eta^2 = 0.008$; PBI: $F(1,9) = 2.359$, $p = 0.159$, partial $\eta^2 = 0.208$). In other words, there was evidence that interacting with Recorded mode before Takeover mode strengthens the social presence perceived by participants for real-human controlled agent avatars compared with the one perceived for recorded agent avatars.

VR Experience A two-way mixed ANOVA test showed there was no significant interaction between whether having VR experience before or not and the mode of interaction on CP ($F(1,18) = 0.017$, $p = 0.898$, partial $\eta^2 = 0.001$) and PBI score ($F(1,18) = 0.008$, $p = 0.928$, partial $\eta^2 < 0.001$). No significant difference was found in mean CP ($F(1,18) = 2.147$, $p = 0.160$, partial $\eta^2 = 0.107$)

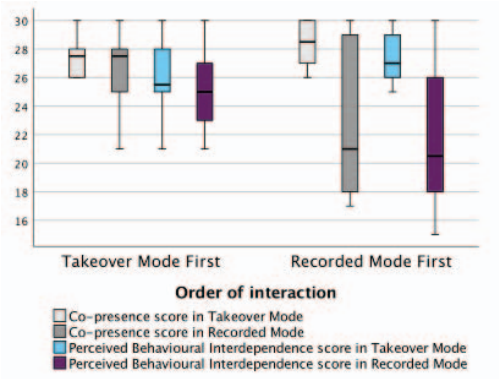

Figure 7: Social presence score in each section for each interaction mode with different interaction order.

and PBI ($F(1,18) = 1.830$, $p = 0.193$, partial $\eta^2 = 0.092$) score between having VR experience or not.

5.2 Interview

5.2.1 Number of Real-human Controlled Avatars Reported

During the interview, the average number of real human controlled avatars reported by the participants for the Takeover Mode ($M = 1.500$, $SD = 1.192$) was higher than the Recorded Mode ($M = 0.750$, $SD = 0.967$). The takeover system in Takeover mode elicited a significant increase in the number of real-human controlled avatars reported by the participants compared to the Recorded mode ($M = 0.750$, 95%CI [0.108, 1.392], $t(19) = 2.445$, $p = 0.012$). Therefore, our hypothesis H2 about the perceived number of real human controlled avatars can be accepted.

A two-way mixed ANOVA test found that changes in the reported number of real-human controlled avatar perceived in different mode varied for different interaction orders ($F(1,18) = 18.356$, $p < 0.001$, partial $\eta^2 = 0.505$).

The number reported in Takeover mode was significantly greater when interacting with Recorded mode first compared to ($M = 1.400$, $SE = 0.437$, $p = 0.005$) with Takeover mode first. However, the effect of order was not found in Recorded mode ($F(1,18) = 1.364$, $p = 0.258$, partial $\eta^2 = 0.070$).

When interacting with Recorded mode first, the number of real-human controlled avatars reported was significantly larger in Takeover mode compared to Recorded mode ($M = 1.700$, $SE = 0.300$, $p < 0.001$). However, the effect of mode was not significant when interacting with Takeover mode first ($F(1,9) = 0.375$, $p = 0.555$, partial $\eta^2 = 0.040$). In other words, there was evidence showing that interacting with Recorded mode before Takeover mode enhanced the perceived feeling of avatars being controlled by a real human in Takeover mode than the one in Recorded mode.

5.2.2 Takeover process

In the interviews, most participants reported taking the three agent avatars as individual entities and did not have the feeling of them being controlled by the same person. Some reported that they were not sure because the agent avatars did not interact with each other and they did not interact with the participants together.

When the functioning of the Takeover mode was explained, most participants were surprised by the fact that the agent avatar was taken over in turn. None of the participants reported noticing the takeover process (i.e. unnatural movements or jerky animation that might have made taking over or releasing the agent avatar noticeable).

5.2.3 Criteria of determining human-controlled or agent

The interview also inquired about what made the participants think the avatar is real-human controlled or not, and what makes them feel suspicious.

Ability to perform unexpected actions The unpredictability of an avatar's actions was usually taken as a criterion for judging whether it was controlled by a human. If the avatar's response was simple and predictable, participants might assume there was a predefined trigger condition prompting the avatar's response. According to the participants, if they "give a response conventionally, it is not a surprise the avatars will also give it back". Such reactions could easily be programmed with triggers based on the distance or sound. In particular, they suggested that the avatar sitting in front of the laptop responded to the user's help in adding milk by clapping his hands, more as feedback to the environment than as direct interaction with the user.

Enthusiastic avatars were more likely to surprise participants. The indifferent response of the avatar might have convinced participants that they were not being controlled by a human. " They seem to live in their own world", as participants said, "they are like NPCs, they don't do anything extra, they don't surprise me in any way". Participants often mentioned that they judged whether a human was in control based on the level of enthusiasm when the avatar greeted them. "He wants to shake my hand which is an unexpected action", said the participant, "you can feel they are trying to react with you".

Sometimes, however, being too enthusiastic conversely led participants to believe the avatar was not controlled by a human. One participant reported that when the barista in the Recorded Mode simply pointed to the doughnuts indifferently, it was more realistic. In contrast, the barista in the Takeover mode greeted warmly before guiding to the doughnuts might be less likely to happen in real life.

Conventional response The conventional response of the avatar could lead the participants to believe it is more likely being controlled by a human. "When I didn't find out where the doughnut was, he pointed further that way and got closer to the doughnut", this reason made several participants perceive the takeover mode to be realistic. If the participant was accidentally very close to the avatar when saying "hi" to them and the avatar in Recorded Mode did not step back to make more space, it was considered not to be controlled by a human.

When participants interfered with the avatar, the avatar was supposed to express dissatisfaction. For example, if the participant deliberately poured milk on the laptop, the avatar was considered realistic if it tried to stop the participant. The avatar in Recorded Mode continued to respond by clapping its hands in this situation which was considered unrealistic.

However, conventional actions sometimes enhanced the feeling of being programmed with a trigger condition. One participant reported that it was "too reasonable for the avatar to clap for me when the glass was almost full", and conversely, "when I accidentally poured milk on the table, the avatar clapped for me as if it was mocking me, making things feel more real."

Size of movement area A larger range of avatars' movement area might give the impression that they were controlled by humans. One participant commented:"The avatar that waved at me in front of the coffee test would walk around the room, while the others would just do repetitive things in their fixed area". Another commented that "The barista exit and enter the bar" made the participant feel this avatar was more realistic.

This result depended on whether the participant was paying attention to the avatar's movement before interacted with it. Although the recorded motion clip being played was long enough so the participants would theoretically not see any repeated motion, they tended to think that the avatar moving in a small area was doing something

in repetition, especially the one sitting in front of the laptop without moving a lot.

Facing direction Whether the avatar was facing the participant when responding was an important criterion. Sometimes due to the participants' lack of proficiency they did not walk directly toward the avatar but were teleported to the place between the avatar and the coffee test board. The Recorded Mode was recorded with an action of turning around and then greeting the participant, so in this mode, the avatar might start waving without facing the participants because they were not in the expected position.

Response speed Slight delays or too quick responses could give the impression that the avatar was not controlled by a human. If a participant tried to interact with the avatar and the avatar did not react in time or reacted too quickly, for example, because it was not taken over by the actor in time, or the actor knew that the participant was going to greet them and turned around early, then this reduced the likelihood that participants would think that this avatar was controlled by a human. In Recorded Mode, it was difficult to determine when it was appropriate to start playing the clip, e.g., when the clip was played after the participant had started to greet them, because the avatar needed to turn around before the greeting and each participant started in a different position.

Avatar's motion Some participants found the movements and interactions of the avatars to be very mechanical. "Turning heads and waving hands in a mechanical, stereotypical, frequency-based movement", "It feels like the movements are programmed" as said by the participants. It was difficult for them to tell why they feel this way, but they still said that this was the reason they believed that avatars were not controlled by humans. We note that the animations were all recorded from a real person acting as described in Sect. 4.1.

6 Discussion

6.1 Noticing the Takeover Process

None of the participants reported they noticed the takeover process. However, this might due to the fact that participants' attention was only on the avatar when they were interacting with it. Therefore, if the avatar was taken over before the participants interacted with it (which is how we usually did in the experiments), they might not observed the process carefully and therefore did not see it. This suggests that the takeover system might be extended with an assessment of visibility and proximity to the agent avatar being taken over.

6.2 Influence of Tasks

In order not to reveal the content of the experiment, participants were told that they needed to complete the task and were not told that they needed to pay attention to the avatar's interaction with them. Therefore, some people may focused entirely on the task and made no effort to interact with the avatar. "I didn't give them the real-life response that I used to give", said the participant after being told about the aim of the experiment.

Some relatively complex tasks such as pouring milk might resulted in a failure to perceive the avatar's clapping in both rounds, leading the participants to believe the avatar was not interacting with them at all and was therefore not in control of a human. Interestingly, noticing the avatar clapping in appreciation may even cause participants to think it is controlled by a human when it is not.

6.3 Within-subject design

Since we designed a within-subject experiment to augment the dataset, participants would enter the same scene twice to experience different conditions. Although they were told to take the two rounds separately, they would still unconsciously provide the answers after comparison.

When the participants first entered the scene, they might be more nervous and therefore tend to focus more on the environment, their tasks and the operation of the controller, rather than what the avatar they are interacting with is doing. ".. is more realistic because the avatar is clapping back to me", said by the participants, who was surprised after been told both round the avatar clapped for their help. Additionally, the first time participants entered the scene, they might be more excited or even attempt something out of the frame. "I'm going to try if avatar will get mad if I pour the milk on its computer" said by the participant in the first round who finished the tasks in the second round relatively fast. This might cause them to miss some of the details of the second round. Conversely, one could argue that participant might be more attentive to the avatars and tasks the second time. In either case, we note that participants did not notice the takeover process, and thus the prototype was successful.

7 Conclusion and Future Works

In this paper, we presented a system to enable a single actor to jump between several avatars and inhabit them in a virtual cafe scenario. The actor can smoothly take control of different avatars by interrupting pre-recorded loopable clips.

A subsequent experiment analyzed whether the system creates believable interactive virtual environments. The within-subject experiment let the two groups of participants interact with both agent avatars in Takeover Mode which has a real actor take control of the avatar when interacting and Recorded Mode in which avatars respond with pre-recorded clips. The result of this experiment shows that the takeover system succeeded in making most participants unaware of the takeover process and the fact that only one actor was playing three roles at the same time. The Takeover Mode gives the participants a better perception of social presence than Recorded Mode and the average number of real human controlled avatars reported is larger in Takeover Mode as well.

Given the apparent success of the prototype, the takeover system still has some limitations that could be improved in future work. While only one actor was involved in the experiment to maintain consistency, an informal lab test showed that other users could learn the takeover process in a few minutes. However, the manner in which takeover is achieved can certainly be improved. The visualisation of future action might be improved, or novel interpolation techniques that modify the pre-recorded animation to the actor. This could assist the actor in choosing the next takeover avatars and in turn poses the opportunity to let the participants have more freedom in the experiment without the need to follow a specific task sequence. Additionally, adapting pre-recorded clips to accommodate variations in actors' height and arm length could make the system applicable to various actors and being used in broader scenarios. To further enhance the system, an interaction score can be implemented to reflect the participant-avatar interaction probability. We only dealt with three point tracking. It may be harder for an actor to mimic full body tracking, but we think this is achievable though it might need more training for the actor. This prototype did not address voice acting, which is an important future step. To keep the participants from noticing the agent avatars are controlled by the same actor, voice transformation, which is now well-developed [32], could be implemented.

Overall our prototype demonstrates that an actor is able to control more than one avatar in a scene. This can have applications directly in small scale systems, such as the experimental scenarios that are commonly used to evaluate social VR itself. With further refinement, it could be a generally useful technique for acting in larger scale experiences.

REFERENCES

[1] F. Anderson, T. Grossman, J. Matejka, and G. Fitzmaurice. YouMove: Enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, pp. 311–320. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2501988.2502045

[2] J. Bailenson and J. Blascovich. Avatars. *Encyclopedia of human-computer interaction*, pp. 64–68, 2004. Berkshire Publishing Group.

[3] F. Biocca, C. Harms, and J. K. Burgoon. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments*, 12(5):456–480, Oct 2003. doi: 10.1162/105474603322761270

[4] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: The behavior expression animation toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 477–486. Association for Computing Machinery, New York, NY, USA, 2001. doi: 10.1145/383259.383315

[5] J. C. Chan, H. Leung, J. K. Tang, and T. Komura. A virtual reality dance training system using motion capture technology. *IEEE Transactions on Learning Technologies*, 4(2):187–195, Apr–Jun 2011. doi: 10.1109/TLT.2010.27

[6] P. T. Chua, R. Crivella, B. Daly, N. Hu, R. Schaaf, D. Ventura, T. Camill, J. Hodgins, and R. Pausch. Training for physical tasks in virtual environments: Tai Chi. In *Proceedings of IEEE Virtual Reality*, pp. 87–94, 2003. doi: 10.1109/VR.2003.1191125

[7] C. Curtis, S. O. Adalgeirsson, H. S. Ciurdar, P. McDermott, J. Velásquez, W. B. Knox, A. Martinez, D. Gaztelumendi, N. A. Goussies, T. Liu, and P. Nandy. Toward believable acting for autonomous animated characters. In *Proceedings of the 15th ACM SIGGRAPH Conference on Motion, Interaction and Games*. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3561975.3562941

[8] S. Dermouche and C. Pelachaud. Generative model of agent's behaviors in human-agent interaction. In *International Conference on Multimodal Interaction*, pp. 375–384. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3340555.3353758

[9] G. C. Dobre, M. Gillies, and X. Pan. Immersive machine learning for social attitude detection in virtual reality narrative games. *Virtual Reality*, 26:1519–1538, Apr 2022. doi: 10.1007/s10055-022-00644-4

[10] Z. Ersozlu, S. Ledger, A. Ersozlu, F. Mayne, and H. Wildy. Mixed-reality learning environments in teacher education: An analysis of TeachLivE™ research. *SAGE Open*, 11(3), 2021. doi: 10.1177/21582440211032155

[11] W. Feng, A. Kannan, G. Gkioxari, and C. L. Zitnick. Learn2Smile: Learning non-verbal interaction through observation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4131–4138, 2017. doi: 10.1109/IROS.2017.8206272

[12] Y. Ferstl and R. Mcdonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 93–98. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3267851.3267898

[13] S. J. Friston, B. J. Congdon, D. Swapp, L. Izzouzi, K. Brandstätter, D. Archer, O. Olkkonen, F. J. Thiel, and A. Steed. Ubiq: A system to build flexible social virtual reality experiences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3489849.3489871

[14] M. Gillies. Creating virtual characters. In *Proceedings of the 5th International Conference on Movement and Computing*. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3212721.3212835

[15] G. Gomes, C. A. Vidal, J. B. Cavalcante Neto, and Y. L. B. Nogueira. An emotional virtual character: A deep learning approach with reinforcement learning. In *21st Symposium on Virtual and Augmented Reality (SVR)*, pp. 223–231, 2019. doi: 10.1109/SVR.2019.00047

[16] N. Hamanishi, T. Miyaki, and J. Rekimoto. Assisting viewpoint to understand own posture as an avatar in-situation. In *Proceedings of the 5th International ACM In-Cooperation HCI and UX Conference*, pp. 1–8. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3328243.3328244

[17] C. Harms and F. Biocca. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, 2004.

[18] T. N. Hoang, M. Reinoso, F. Vetere, and E. Tanin. Onebody: Remote posture guidance system using first person view in virtual environment. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2971485.2971521

[19] C. Kyrlitsias and D. Michael-Grigoriou. Social interaction with agents and avatars in immersive virtual environments: A survey. *Front. Virtual Real*, 2, Jan 2022. doi: 10.3389/frvir.2021.786665

[20] Laerd Statistics. Two-way mixed anova using spss statistics. statistical tutorials and software guides. 2015. Retrieved from: https://statistics.laerd.com/.

[21] S. P. Lee, J. B. Badler, and N. I. Badler. Eyes alive. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 637–644. Association for Computing Machinery, New York, NY, USA, 2002. doi: 10.1145/566570.566629

[22] J. Liu, Y. Zheng, K. Wang, Y. Bian, W. Gai, and D. Gao. A real-time interactive Tai Chi learning system based on VR and motion capture technology. *Procedia Computer Science*, 174:712–719, 2020. doi: 10.1016/j.procs.2020.06.147

[23] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. 2023. doi: 10.48550/arXiv.2304.01852

[24] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 25–35. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2485895.2485900

[25] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph.*, 27(1):1–24, mar 2008. doi: 10.1145/1330511.1330516

[26] X. Pan and A. Hamilton. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, 109(3):395–417, Mar 2018. doi: 10.1111/bjop.12290

[27] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. 2023. doi: 10.48550/arXiv.2304.03442

[28] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:3549–3557, Dec 2009. doi: 10.1098/rstb.2009.0138

[29] P. J. H. Slijkhuis. *The Uncanny Valley Phenomenon: A replication with short exposure times*. PhD thesis, University of Twente, 2017. doi: 10.13140/RG.2.2.36658.73927

[30] A. Steed, L. Izzouzi, K. Brandstätter, S. Friston, B. Congdon, O. Olkkonen, D. Giunchi, N. Numan, and D. Swapp. Ubiq-exp: A toolkit to build and run remote and distributed mixed reality experiments. *Frontiers in Virtual Reality*, 3, Oct 2022. doi: 10.3389/frvir.2022.912078

[31] N. Stephenson. *The Diamond Age*. Bantam Books, 1995.

[32] Y. Stylianou. Voice transformation: A survey. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3585–3588, 2009. doi: 10.1109/ICASSP.2009.4960401

[33] A. M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, Oct 1950. doi: 10.1093/mind/LIX.236.433

[34] H. Warpefelt. *The Non-Player Character: Exploring the believability of NPC presentation and behavior*. PhD thesis, Stockholm University, 2016.

[35] S. Yang, T. Li, X. Gong, B. Peng, and J. Hu. A review on crowd simulation and modeling. *Graphical Models*, 111, Sep 2020. doi: 10.1016/j.gmod.2020.101081

[36] T. Yin, L. Hoyet, M. Christie, M.-P. Cani, and J. Pettré. The One-Man-Crowd: Single user generation of crowd motions using virtual reality. *IEEE Transactions on Visualization and Computer Graphics*,

28(5):2245–2255, 2022. doi: 10.1109/TVCG.2022.3150507
[37] C. Youngblut. Experience of presence in virtual environments. Sep 2003. doi: 10.21236/ada427495
[38] X. Zhang, E. Wu, and H. Koike. Watch-your-skiing: Visualizations for VR skiing using real-time body tracking. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 387–388, 2021. doi: 10.1109/ISMAR-Adjunct54149.2021.00088