Not Alone Here?! Scalability and User Experience of Embodied Ambient Crowds in Distributed Social Virtual Reality

Marc Erich Latoschik, Florian Kern, Jan-Philipp Stauffert, Andrea Bartl, Mario Botsch, and Jean-Luc Lugrin

Fig. 1. An immersive Social Virtual Reality (SVR) with multiple avatars and agents co-located in the same virtual space as seen from an immersed participant's point of view. Our SVR system supports large crowds of distributed avatar/agent participants of variable appearances (see an abstract virtual body in the front left, and a photorealistic virtual body from photogrammetry scans on the right).

Abstract—This article investigates performance and user experience in Social Virtual Reality (SVR) targeting distributed, embodied, and immersive, face-to-face encounters. We demonstrate the close relationship between scalability, reproduction accuracy, and the resulting performance characteristics, as well as the impact of these characteristics on users co-located with larger groups of embodied virtual others. System scalability provides a variable number of co-located avatars and AI-controlled agents with a variety of different appearances, including realistic-looking virtual humans generated from photogrammetry scans. The article reports on how to meet the requirements of embodied SVR with today's technical off-the-shelf solutions and what to expect regarding features, performance, and potential limitations. Special care has been taken to achieve low latencies and sufficient frame rates necessary for reliable communication of embodied social signals. We propose a hybrid evaluation approach which coherently relates results from technical benchmarks to subjective ratings and which confirms required performance characteristics for the target scenario of larger distributed groups. A user-study reveals positive effects of an increasing number of co-located social companions on the quality of experience of virtual worlds, i.e., on presence, possibility of interaction, and co-presence. It also shows that variety in avatar/agent appearance might increase eeriness but might also stimulate an increased interest of participants about the environment.

Index Terms—Social Virtual Reality, quality of experience, performance characteristics, co-location, co-presence, possibility of interaction, ambient crowds, avatars and agents, computer-mediated communication, multi-user virtual environment.

1 Introduction

Embodied Social Virtual Reality (SVR) exploits the rich social signals and behavior patterns humans use in the physical world [44]. These signals significantly originate from our paraverbal and non-verbal expressions in face-to-face encounters. Body movements, gestures, mimics, and eye movements play crucial roles in social behavioral phenomena like joint attention, grouping, eye contact, or mutual synchronization and coordination [38]. Embodiment technologies provide the necessary means to realize virtual face-to-face encounters enabling social signals via so-called avatars, our digital alter egos in the virtual realm.

- *Marc Erich Latoschik, Florian Kern, Jan-Philipp Stauffert, Andrea Bartl, and Jean-Luc Lugrin are with the HCI group of the University of Würzburg. E-mail: marc.latoschik@uni-wuerzburg.de.*
- *Mario Botsch is with Bielefeld University. E-mail: botsch@techfak.uni-bielefeld.de.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

SVRs have gained much interest in both, academia and industry. Companies like Second Life[1], AltspaceVR[2], Pixo VR[3], or NVIDIA with its Holodeck[4] project already developed real-world applications or impressive demonstrations. Although, to some extent, these developments suggest a solid maturation of the overall field, in fact, the existing approaches differ in significant aspects. Schroeder [40] separates such Multi-User Virtual Environments (MUVEs) into *immersive environments*, e.g., NVIDIA's Holodeck project, and *online worlds*, e.g., Second Life or typical networked multi-user computer games. We argue that this distinction is drawn from the availability or absence of embodiment features and its qualities in terms of (1) completeness of represented and controllable body parts, (2) the avatars' appearances or looks, and (3) direct control of the avatars' bodies with a sufficient sensory coverage of the controlling users' movements in real-time.

Overall, embodied SVR promises novel forms of computer-mediated communication but it is particularly challenging regarding sensory

[1] https://secondlife.com
[2] https://altvr.com
[3] https://pixogroup.com
[4] https://www.nvidia.com/en-us/design-visualization/technologies/holodeck/

coverage and temporal and precision requirements, which is considered one of the grand challenges for VR [41], specifically if it has to be realized with distributed systems. Table 1 derives hypothetical data rates while scaling up sensory coverage and the number of co-located avatars for a selection of fidelities. Typically, setups with just two embodied avatars in 1:1 dyadic social avatar-avatar encounters [4, 5, 25, 39, 44] do not need to utilize distribution. Similarly, work on the effect of larger crowds of virtual humans also either use computer-controlled agents [33] or utilize pre-recorded material presented as videos [2, 9]. Hence, little is known about the experience of users fully immersed and embodied in a simulated Virtual Environment (VE) with larger groups of co-located avatars with variable appearance, and potential performance characteristics as given by a real-world distribution of these companions.

Contribution

This article investigates the current state-of-the-art of distributed Social Virtual Realities with consumer VR technology, and the effects of larger embodied ambient crowds on participating users. The approach deliberately utilizes today's consumer VR technology, i.e., the Unreal Engine 4 (UE4), for two reasons: (1) It targets out-of-lab real-world applications. (2) To show potential benefits and limitations of today's consumer technology when realizing distributed SVRs. A scalable system architecture supports various types and fidelities of control schemes and avatar appearances, up to high-quality, realistic looking individualized avatars captured by photogrammetry. We propose a hybrid evaluation combining technical benchmarks with user-centered tests. We demonstrate the hybrid evaluation for up to 125 participants and show the close relation between objective measures and subjective ratings. The evaluation confirms sufficient performance for 25 participants in a typical real-world distribution scenario. This upper bound is applied to and verified by a user study on the experience of users immersed in an SVR with co-located ambient crowds. To the best of our knowledge, this evaluation reveals two novel effects: (1) An increasing number of co-located social companions has a positive effect on the quality of experience of virtual worlds, i.e., on presence, possibility of interaction, and co-presence. (2) It also shows that variety in avatar/agent appearance might increase eeriness but might also stimulate an increased interest of participants about the environment.

2 Related Work

Schroeder gives a comprehensive overview of the fundamental aspects of social interaction in virtual worlds [40]. Steed and Schroeder also highlight some fundamental concepts as well as technical aspects and requirements for SVRs [47]. They classify current embodiment approaches along a scale representing the degree of user modeling and identify three main types of current approaches: (1) puppeteered, (2) reconstructed, and (3) tracked. Typical online games, as well as multi-user worlds similar to Second Life, are basically puppeteered. Avatars here are controlled by some non-direct animation scheme, e.g., by pressing a button to trigger an animation. With an immersive first-person perspective from inside one's avatar, this control scheme seems detrimental. Still, even puppeteered systems motivate research in SVR [20, 32], specifically about alternative social mechanics.

Real-time reconstruction of dynamic scenes and users promises a faithful dynamic replication of real physical appearances and environments [3, 13, 35]. It certainly is appropriate for many use-cases, e.g., teleconferencing. However, it does not allow to easily modify avatar appearance, as, e.g., is required by work on the *illusion of virtual body ownership (IVBO)* [18, 30, 42] or the *Proteus* effect [51], specifically in avatar-avatar encounters [25] or dyadic social avatar-avatar interactions [4, 5, 37, 39]. A deliberate change of the appearance of avatars could also be desirable to avoid stigmatizing in SVR. Additionally, concerning scalability, dynamic 3D reconstruction is characterized by high bandwidth requirements. These exceed typical requirements of tracked approaches (see Table 1) by orders of magnitudes even with appropriate compression [26], which often will also increase the latency.

Tracked avatar embodiment for SVRs requires direct control schemes of as many degrees of freedom as the human body has, and hence, a) elaborated sensor technology like full-body motion tracking [21, 45] and/or face tracking [24], and b) an appropriate model of a virtual human body matching the sensory input. Such models typically consist of a properly rigged body mesh for skeletal animations together with blend shapes for facial animations where applicable. They are either generated manually via 3D-modeling, or via off-line reconstruction from real humans [1, 10] (or a combination of both), effectively combining dynamic tracking with static reconstruction.

The extent of sensory coverage depends on the reproduction accuracy between the controlling user and the controlled avatar and its body and animation model. A full embodiment of self-avatars increases presence but full as well as partial embodiment increases co-presence [15], although recent work could not substantiate these findings [28], which might be caused by strong contextual distractors. Recently, initially single-user embodiment studies also started to explore the effect of the appearance and the behavior of an other's avatar co-located in SVR [25, 39]. All agree on the necessity of low latencies for contingencies [49], e.g., convincing visuomotor synchrony. Additional requirements are a first-person perspective, a sufficiently realistic avatar appearance [2, 31, 48], and a high degree of immersion [48]. Work which includes larger groups of others so far either used non-distributed settings of up to eight virtual agents with comparable looks [6, 43], or even non-immersive and non-interactive pre-recorded videos [9].

Table 1. Estimations of net data transfer rates required to communicate tracked non-verbal behavior of avatars of different fidelities. Accurate numbers are subject to a variety of design choices and optimizations: (1) applied body model, (2) sensory coverage, (3) model-based optimizations, (4) resolution, i.e., number of bits per value, and (5) compression. Model-based optimizations refer to potential uses of forward kinematics (FK), or inverse kinematics (IK), for shared models. IK is potentially applicable, i.e., for linked models (supporting a proper skeleton with chained joints and links) based on B1-B3, but could also be applied to partial models. FK is necessary for linked models based on B4-B6. Results reflect a coarse upper bound estimation: no potential overhead, no compression applied, single precision float values yielding 4 bytes, representation of **p**(osition) and **r**(otation) each by 3 floats, **f**(lexion) and **a**(bduction) each by one float. For the face we include the seven basic emotions of Ekman and Friesen in F1 and a selection of 44 Action Units of the Facial Action Coding System [11] encoded with either 3 bit (F2) or 1 float (F3).

			Bytes / avatar	**Data rate in KB/s per number of clients (1 avatar/client) at 90 Hz**							
				2	**5**	**10**	**25**	**50**	**75**	**100**	**125**
Fidelity of body model											
B1	head	$1 \times \mathbf{pr}$	24.0	4.3	10.8	21.6	54.0	108.0	162.0	216.0	270.0
B2	+ 2 hands	$3 \times \mathbf{pr}$	72.0	13.0	32.4	64.8	162.0	324.0	486.0	648.0	810.0
B3	+ 2 feet and spine	$6 \times \mathbf{pr}$	144.0	25.9	64.8	129.6	324.0	648.0	972.0	1296.0	1620.0
B4	skeleton medium	$1 \times \mathbf{p} + 17 \times \mathbf{r}$	216.0	38.9	97.2	194.4	486.0	972.0	1458.0	1944.0	2430.0
B5	hand low	$5 \times \mathbf{f} + 5 \times \mathbf{a} + \mathbf{r}$	52.0	9.4	23.4	46.8	117.0	234.0	351.0	468.0	585.0
B6	hand high	$15 \times \mathbf{f} + 5 \times \mathbf{a} + \mathbf{r}$	92.0	16.6	41.4	82.8	207.0	414.0	621.0	828.0	1035.0
Fidelity of face model											
F1	2 eyes + 7 emotions	$2 \times \mathbf{r}$ + 7 float	52.0	9.4	23.4	46.8	117.0	234.0	351.0	468.0	585.0
F2	2 eyes + 44 FACS bit	$2 \times \mathbf{r}$ + 44×3 bit	40.5	7.3	18.2	36.5	91.1	182.3	273.4	364.5	455.6
F3	2 eyes + 44 FACS float	$2 \times \mathbf{r}$ + 44 float	200.0	36.0	90.0	180.0	450.0	900.0	1350.0	1800.0	2250.0

2.1 Discussion and Requirements

We currently have little knowledge about the experience of users co-located with larger groups of avatars realized with distributed immersive and embodied SVR. How does it feel to be surrounded by virtual others? Current game technology provides sophisticated rendering and networking features. SVRs are sensitive to performance characteristics due to the close temporal patterns of non-verbal social signals. How does current technology cope with the extensive embodiment requirements and how do the specific performance characteristics impact user experience regarding scalability? To answer these questions, we chose a hybrid approach combining tracked user motions and avatar models (up to avatars reconstructed by photogrammetry) for three reasons: (1) To support applications requiring scalability in terms of modified avatar appearances, (2) to scale up the number of distributed avatars and/or sensory coverage, and (3) to be compatible with animation principles of current game engines.

Following the theoretical estimates in Table 1, we chose a medium fidelity B3 (324.0 KB/s) for the current evaluation. Note that B3 potentially either requires IK to be in effect at the distributed clients, or it has to do without any linked models at all. However, its resulting bandwidth is roughly comparable to B4, which transmits all data necessary to replicate complete linked models. Hence, B3 is a suitable candidate for our upcoming performance evaluation, which are mainly targeting impacts from potential bandwidth and latency bottlenecks caused by a real-world distribution.

UE4 promotes a multi-user client-server distribution architecture. We assume a standard 1 Gbit network link from the server to the internet backbone, and clients connected with potentially much less bandwidth, e.g., from private homes. A proper multicast infrastructure can in general not be expected for the given distribution scenario. Hence, an increasing number of clients would certainly increase the load on the server, specifically for outgoing replication. For example, given 25 clients and fidelity B3, we require a total bandwidth of 324 KB/s to the server, and 8100 KB/s from the server (who has to replicate all data back to all clients) at 90 Hz. The resulting client bandwidth here is much lower (ca. 13 KB/s up; 324 KB/s down). Overall, the final requirements for the developed system are as follows:

- R1 SVR supporting a scalable number of physically distributed users.
- R2 Real-world application with potentially novice end-users.
- R3 High immersion with first-person perspective.
- R4 Full embodiment with sufficient sensory coverage.
- R5 Variable, realistic avatar appearance.
- R6 High visuomotor synchrony and responsiveness.
  - (a) Sufficient data throughput.
  - (b) Low latencies and low latency jitter.
  - (c) High data fidelity (accuracy and precision).

Similarly to work in [23, 29], our system supports mixed virtual crowds of user-controlled avatars with AI-controlled virtual agents for various application-specific tasks (e.g., as role-models or troublemakers). However, this article does not focus on any agent-specific research questions which are the topic of an alternative publication [27].

Benchmarking the non-functional requirements R6 for VR systems often uses two general approaches in combination. Intrusive benchmarking like for real-time systems, in general, requires instrumentation of the code itself. Elaborate VR frameworks and game engines usually support intrusive benchmarking and provide additional tools for profiling. The intrusive approach also allows pinpointing sources of problems inside the code. On the downside, it requires full-blown code access, measurements potentially interfere with the results, and the outcomes do not necessarily correlate to end-user experiences. Hence, VR-benchmarking often applies non-intrusive black-box benchmarking and end-to-end measurements. Non-intrusive benchmarks include camera-based latency measures by phase shift analysis of sine curve movements [46] or body movements [49], automated frame counting [12], or formal simulation of systems [36]. Chang et al. found interesting system behaviors with a non-intrusive high-speed camera-based approach [8]. They identified sensitivity-precision tradeoffs of the underlying engines, which are of high relevance for our approach. Such tradeoffs often result from internal optimizations of the underlying engine potentially out-of-reach for application developers.

Overall, we will use a combination of intrusive and non-intrusive benchmarks to evaluate the specified non-functional performance requirements concerning the scalability features of the system. We will complement the technical benchmarks with subjective ratings of the user experience concerning the performance characteristics (fluidity, synchrony, annoyment, and simulator sickness) to identify potential correlations between both evaluation methods. Finally, we shed some light onto the subjective effects of being inside an SVR populated by a variable number of co-located avatars with different appearances. Here, we use the system to investigate the resulting user experience based on a selection of adequate factors for evaluating an SVR with co-located avatars, e.g., attractiveness, humanness and eerieness, presence, co-presence, and the possibility of interaction.

3 System Description

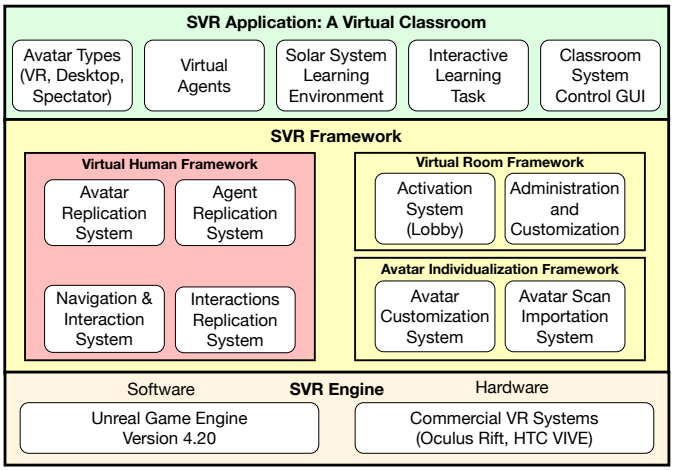

Fig. 2. Overview of the three main layers of the software architecture, which separates specific system functions according to their specificity with respect to the required application domain.

Fig. 2 illustrates the overall software parts of the system organized into three main abstraction layers:

1. SVR Application Layer: Specific functions for customized SVR content, e.g., for distributed embodied learning or training.
2. SVR Framework Layer: Generic functions for distributed embodied SVR that provide the basic avatar and agent representation and animation capabilities and the environment specification.
3. SVR Engine Layer: Underlying hardware and software functions supporting device Input/Output (I/O), basic interaction schemes, core visualization and simulation capabilities, network communication facilities, and software component integration schemes.

The following sections describe the SVR Engine and Framework layers in detail and encompass the relevant functionality for general SVR support. We take a closer look at the distribution and networking architecture which follows common practices proposed by the Unreal Engine development community and builds upon the provided UE4 network replication methods.

3.1 SVR Engine Layer: Hard- and Software

Requirements R2 and R3 are the determining factors for the use of consumer VR hard- and software. End-users have to operate the system from their homes without the help of a technician or trained personnel. To also support R3, the system uses Oculus Rift as well as HTC VIVE and HTC VIVE pro head-mounted displays. UE4 provides application packaging and distribution necessary to support R2. The

engine has proven to be beneficial in related work [23, 29]. It has a well-known reputation for rendering high-quality virtual humans. It is used in combination with the photogrammetry-based method to capture high-quality avatars following [1, 48] to fulfill R5. Elaborated software tools like the UE4 usually already support several essential features, which are necessary to implement the functional requirements. Notably, they often also predefine how to satisfy a specific functionality, and they enforce a particular development model. This guidance is helpful for novices but also restrictive for experienced programmers. Additionally, such tools often incorporate idiosyncratic terminology for their programming primitives, which complicates comprehension of existing correlations to important software engineering concepts. In the following section, we try to pinpoint differences where appropriate, but most often adhere to the provided programming model and the resulting terminology and naming. We deliberately made this choice since one goal of this work is to identify where we stand concerning the realization of SVR with the given technology. It should also foster replicability since this specific terminology is used throughout the UE4 documentation.

3.2 SVR Framework Layer

3.2.1 Virtual Human Framework

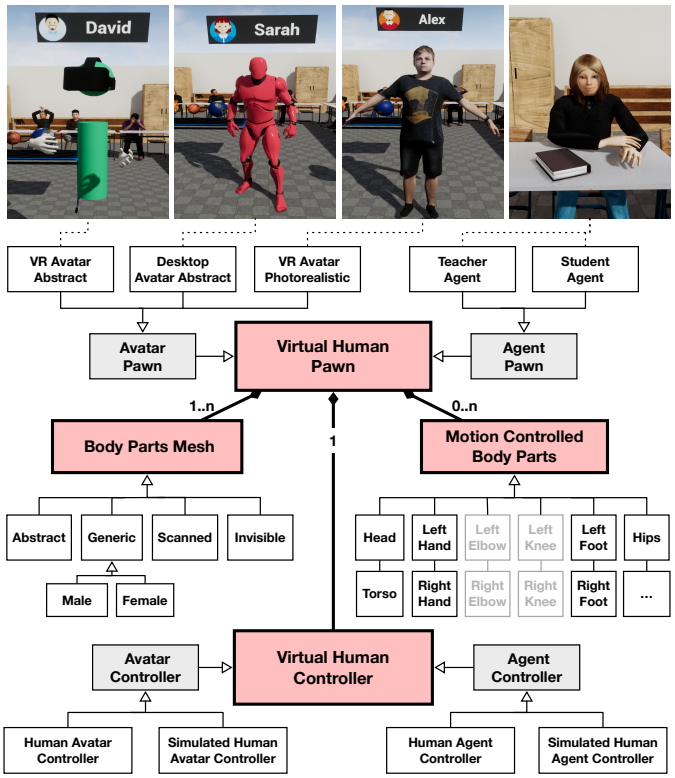

Fig. 3. Class diagram of the Virtual Human Framework (bottom). It supports variable body models (top) in accordance with the *Pawn* and *Controller* abstractions proposed by UE4.

Fig. 3 illustrates the architecture of the *Virtual Human Framework*, which collectively supports virtual humans controlled by a user (i.e., avatar) as well as controlled by the system (i.e., agent). The framework models each virtual human as *Virtual Human Pawn*, which is composed of a *Virtual Human Controller* and a combination of *Body Part Mesh* and *Motion Controlled Body Part*. A flexible combination of different body models and motion-controlled body parts provides various specializations of *Virtual Human Pawn*, such as our *VR Abstract Avatar*, *VR Photorealistic Avatar*, or the *Desktop Avatar*. The *Controller* classes contain the logic responsible for driving the animations of the pawn. Any pawn can be controlled in real-time either (1) puppeteered using users' keyboard, mouse, or controller inputs, (2) tracked using sensory input of users' movements and expressions, or (3) algorithmically animated based on artificial intelligence techniques (e.g., behavior tree or scripted scenario). The classes *Simulated Human Avatar Controller* and *Simulated Human Agent Controller* provide valuable features for testing and benchmarking. They provide pre-recorded tracking data or random movement sequences for any pawn type. They also permit to control the input frequency and amplitudes in order to create more controllable and realistic variations during benchmarking.

3.2.2 Virtual Room Framework

A lobby menu provides an application's starting point. The lobby supports the selection and administration of the target virtual environment and the user's configuration. The menu allows customizing the avatar of the user: name, image, color, and types (e.g., VR Abstract, Photorealistic, or Desktop). By default, the user who creates the server instance is the administrator. She/He can ban other users and select different types of virtual rooms. Shortcut buttons provide a faster setup of multiple parameters like in the benchmark scenario and allow the administrator to set the user's configuration to default values.

3.2.3 Avatar Individualization Framework

Fig. 4 shows the photogrammetry rig we use to generate photorealistic and individualized avatars. It includes 106 Canon DSLR cameras, model EOS1300D. 96 cameras focus the body, 10 cameras focus the face. The 3D model is generated with the photogrammetry software CapturingReality, and post-processed and cleaned with Autodesk Mudbox. Retopology and polycount reduction, as well as UV mapping, is achieved with R3dS Wrap. The resulting avatars have a polycount of around 40k triangles. For comparison, the standard UE4 mannequin has a polycount of 41k triangles. The UV-mapped textures exported from R3dS Wrap have a resolution of 4096×4096. We use Maya to rig our avatars and to export the results as an FBX file into UE4. Current work integrates pre-processing speed-ups for the avatar models as motivated by [1].

3.3 Distribution Architecture

The distribution architecture and multi-user support follow a client-server model as promoted by UE4. Clients can control avatars as well as agents. The *Virtual Human Pawn* class (see Fig. 3) centralizes the replication semantics per virtual human (avatar or agent) as illustrated in Fig. 5 for two clients. The client packs the updated state of all body parts (e.g., head, hands, and feet) into an array for efficiency, and sends it in bulk to the server via remote procedure calls (RPCs). The server locally stores and replicates this data to all clients. Receiving clients apply the replicated data to the corresponding body parts.

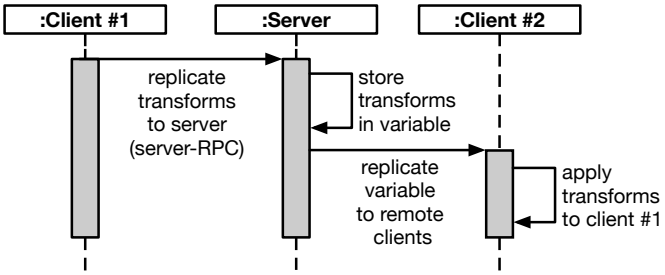

Fig. 5. A client replicates the movements of body parts by executing an RPC on the server. The server replicates this data to the remote client. The remote client applies the replicated movement data to body parts.

UE4 provides several options to realize and parameterize networking. Our current system uses unreliable client-server communication and server-client replication without notification, both to reduce potential latencies. A pre-study did not reveal any significant drop-outs in a comparison of unreliable to reliable client-server communication. We set the replication rate for client-server communication to 60 Hz (desktop clients) and 90 Hz (VR clients) to limit the maximum data

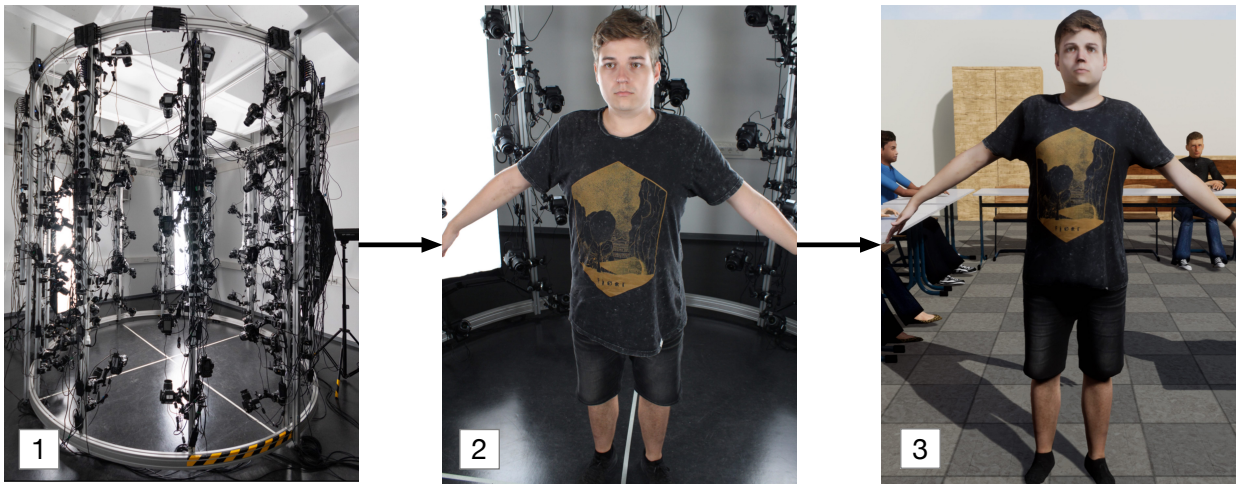

Fig. 4. The photogrammetry rig used to scan individuals (1), a person during the scan (2), and the resulting high quality avatar (3).

rate while still providing smooth animations. Such technical replication parameters will certainly affect the overall characteristics of the system performance and potentially will, in consequence, also affect user-perception in some unpredicted way [8]. Hence, we test the system concerning objective and subjective measures in combination.

4 Evaluation Design, Scenario, and System

The evaluation consisted of three consecutive evaluation phases (EPs):

EP1 **Performance Benchmarking** tested the scalability concerning objective performance characteristics of latency, frame rate, and data rate with an increasing number of clients (2, 5, 10, 25, 50, 75, 100, 125), i.e., avatars participating in the SVR. EP1 recorded the stimulus material for the upcoming evaluation phases: video recordings of the two screens (see Fig. 6) for EP2 and movement recordings, i.e., position and rotation for chosen fidelity, of the collaborating partner and the three planets with an extended movement sequence for EP3. EP1 finally identified 25 as a first potential upper bound for the number of avatars for EP2 and EP3.

EP2 **Performance Perception** tested the scalability concerning the subjective impact of the technical performance characteristics measured in EP1 on perceived fluidity, synchrony, and annoyment on a non-immersed user. EP2 used the video recordings from EP1 with the same scaling conditions. This phase validated and confirmed the upper bound for the number of avatars for EP3.

EP3 **Subjective Experience of Co-Location and Scalability** tested the subjective user experience inside a distributed ambient crowd and the impact of the technical performance characteristics of the recorded movements of the interactions from EP1 with an increasing number of simulated avatars of different avatar appearances (uniformly human-like as in Fig. 3, upper right, and mixed human-like and abstract as in Fig. 3, upper right and left) on an immersed user. EP3 used a reduced number of avatars (2, 10, 25, 100), with the condition 100 explicitly exceeding 25 as the target maximum number of avatars, and confirmed this maximum. These numbers are reported here already as a lookahead to some of the results from EP1 and EP2 for clarity.

The consecutive phases EP2 and EP3 deliberately used prerecorded animations to not induce any confounds by changed stimuli throughout the experiments and to ensure comparability. However, these recordings retained all visibly perceivable performance characteristics resulting from an increased number of clients and avatars, i.e., latencies and stuttering of the interactive animations. The virtual environment for all evaluation phases resembled a classroom with a teamwork-oriented layout, with the participating avatars seemingly collaborating in distributed groups around tables (see Fig. 1). The user and one participant who apparently was directly interacting with her/him are seated vis-a-vis at one table. A virtual solar system is visualized floating in the air in-between them. All executables for the three phases were initially developed using the blueprint visual scripting system of UE4. Phases EP1 and EP2 were nativized, i.e., automatically translated to C++ and compiled as stand-alone executables. This approach is in line with our initial assumption to show what we can expect from today's consumer systems without any further close-to-metal optimizations. Only the recording and playback functions were natively implemented in C++ to reduce any performance overhead from these intrusive functions. The server application in EP3 was not nativized due to an incompatibility with a required plugin but did not require any of the potentially performance-critical networking capabilities.

Table 2. Specifications of the hardware used during the study.

Computers	**CPU**	**RAM**	**GPU**
1× Server	i7-8700K	16GB	NVIDIA GTX 1080 Ti
2× VR Clients	i7-8700K	16GB	NVIDIA GTX 1080 Ti
Load Test Clients			
5× Computers	i7-8700k	16GB	NVIDIA GTX 1080 Ti
9× Computers	i7-7700k	16GB	NVIDIA GTX 1080
18× Computers	i5-6600	16GB	NVIDIA GTX 1080

All phases were implemented using the hardware specified in Table 2. Hosts used 1 Gbit ethernet connected via a switch infrastructure. The VR clients used Oculus Rift HMDs with Oculus Touch controllers. The system was implemented using the Unreal Engine 4.20 and Microsoft Windows 10. Up to 4 instances of the client systems had to share one of the load test hosts for scaling conditions beyond 25 live clients. We took care to distribute client systems to the load test hosts uniformly and to reduce performance impact by the graphics stages as much as possible. Still, multiple clients per host potentially result in additional bottlenecks. However, all results identified to satisfy R1 are not affected by this. The distribution of client and server systems on the available hosts and the control of the avatars and agents were as follows:

EP1: 1 server per server host; 2 VR clients, each with dedicated VR host; uniform distribution of load test clients to the load test hosts following the required scalability conditions. The non-interacting clients simulate simple avatar movements which do not stress the clients but which produce the appropriate data rates.

EP2: No live systems needed. Initially, the setup is the same as for EP1 since it is a recorded video of all animations for the given scalability conditions from EP1.

EP3: One server per server host to play back the recordings of the interaction, and to integrate the participant's avatar inspecting

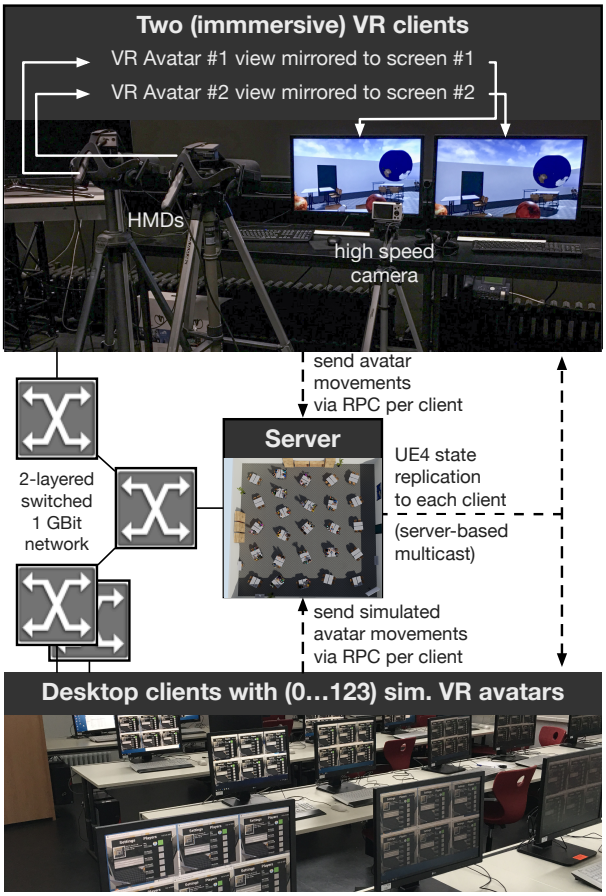

Fig. 6. Physical layout, interconnect scheme (mid left), and logical distribution architecture (mid right) of the benchmarking scenario. On 32 computers, up to four application instances are running for load tests.

the scene. Notably, the play-back did not cause any significant additional load, and the recordings of the movement data of the interactions retained all visually perceivable performance characteristics generated by EP1.

5 EP1 – Performance Benchmarking

Fig. 6 illustrates the physical setup for the performance benchmarking. The displayed scene of the two VR clients is mirrored to the two monitors with a refresh rate of 60 Hz. A high-speed camera is placed in front of the two VR clients to record both screens at 240 Hz. One user continuously performs a smooth drag-and-drop operation of a planet from left to right and back.

5.1 Measuring Latency by Frame Counting

We conducted manual frame counting on the high-speed video following [14] for all scaling conditions (2, 5, 10, 25, 50, 75, 100, 125). We counted how many frames passed between observing updates of the object's location between the two VR clients. Movement smoothing techniques were disabled to see the raw updates. Additionally, we measured the time elapsed between an initiated grab of an object and the reception of this event by the other VR client.

Execution of the triggered event of grabbing a planet completed within a mean of at most 12 ms between the two clients for all test conditions. This delay is equal or below the screen refresh rate of the benchmark monitors (60 Hz $\approx$ 16.6 ms), hence clients receive updates with a delay that is less than the smallest measurable unit. The delay between position updates of an initially smooth movement determines how choppy the movement looks to users. This value increases with an increasing number of connected clients. Fig. 7 visualizes the results of this measurement. Table 3 shows the averaged numbers for all frame

Table 3. Latency as determined by counting how much time passes between initiating a movement on one computer and seeing the movement on a network-connected computer's screen (left) and between two updates of a moving object (right). Higher latency means bigger and more discernable jumps in the movement.

Number of Clients	Latency in ms as means (SD) for Movement Begin	Update Rate
2	8(3.65)	18(06.26)
5	7(2.81)	21(08.94)
10	8(2.64)	19(06.99)
25	8(2.64)	30(06.67)
50	12(7.3)	40(16.60)
75	8(1.96)	57(13.22)
100	8(0)	95(14.82)
125	10(1.3)	158(10.65)

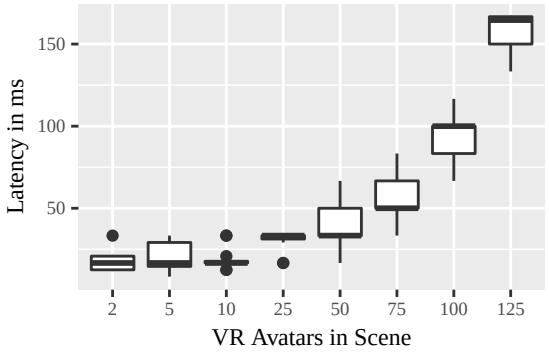

Fig. 7. Average latency and standard deviation between two updates of a movement for a varying number of clients/avatars connected.

counting measurements. As can be seen, there is a notable latency increase detectable beyond 25 clients.

5.2 Measuring Performance by Network Statistics

We measured network performance on the server using the *Stat Net* command and the *Network Profiler* of UE4 for the scaling conditions (2, 5, 10, 25, 50, 75, 100, 125) to find out, whether and how an increasing number of clients and avatars increases the latency and impacts the frame rate. Table 4 reports the results for the benchmarking scenario, i.e., the FPS for server and VR client, the VR client ping for the latency, the network I/O rates on the server and on average per client, the latter four ordered in the sequence of the replication.

The FPS on server and clients decrease with more than 25 connected clients. Also, the clients' latencies increase with an increasing number of connected clients. First, this confirms our results from the frame counting in Sect. 5.1. Second, the results for the data rate from the clients to the server (before the performance drop) are in general estimated by Table 1 when scaled to a 60 Hz replication rate. Observable differences are caused by an enabled compression and a detected basic load of UE4's network layer, which is in effect even without additional payload. Third, the measurements for conditions with 50+ clients illustrate typical challenges in the context of black-box testing, i.e., to precisely pinpoint the source of the bottleneck. With 50+ clients most data rates decrease, or slow-down their increase as for the server out-rate. This certainly is due to the decreased FPS at the server and the clients, since replication between the hosts is bound to the simulation rates. Still, measured data rates are well inside the available bandwidth, rendering a network bottleneck unlikely. Inspecting CPU system performance revealed a high load on one core of the server for the critical conditions. Since the simulation loop is responsible for all replication I/O, our current analysis strongly suggests the performance bottleneck likely to be located somewhere within the server's network I/O capabilities.

Table 4. The server and client performance and network statistics during the benchmarking scenario. The results visualize a decreasing number of frames per second (FPS) at the server and an increasing latency of the clients (Client Ping) with an increasing number of connected clients for measurements beyond 25 clients. Client FPS and data rates decrease slightly time-delayed. See text for further discussions.

Number of Clients	Server FPS	VR Client FPS	Client Ping (ms)	Out Rate Client Avg. (KB/s)	⟶ In Rate Server (KB/s)	⟶ Out Rate Server (KB/s)	⟶ In Rate Client Avg. (KB/s)
2	120	90	9	11.5	23.0	21.5	10.8
5	120	90	9	11.0	54.5	146.9	29.4
10	120	90	11	10.8	107.5	571.8	57.3
25	120	90	13	11.6	288.7	3653.7	143.1
50	40	90	35	10.5	534.7	7081.1	136.4
75	17	60	65	6.4	474.8	9377.3	124.4
100	11	60	117	4.2	1038.8	10163.3	98.7
125	6	35	182	2.8	342.7	11386.2	80.9

Table 5. Descriptive statistics for the three items. For fluidity and synchrony, high values mean high approval. For annoyment low values mean low annoyment. Scales range from 1 to 5.

Number of Clients	Fluidity $M(SD)$	Synchrony $M(SD)$	Annoyment $M(SD)$
2	3.88(1.21)	4.48(.67)	1.67(.82)
5	3.95(1.23)	4.40(.83)	1.57(.97)
10	3.90(1.08)	4.52(.67)	1.57(.86)
25	3.88(1.02)	4.19(.97)	1.93(.92)
50	2.90(1.28)	3.88(.89)	2.48(1.27)
75	2.88(1.11)	3.12(1.13)	2.76(.91)
100	1.81(.97)	2.12(.94)	3.88(.94)
125	1.33(.72)	1.48(.74)	4.64(.58)

6 EP2 – Performance Perception

We recorded eight short video clips of 30 seconds of the interaction described in Sect. 5 for each scaling condition from 2 to 125 clients as before. These videos were provided via an online survey to collect the user feedback about the perceived latency. Each participant watched all eight videos. The order of the videos was randomized. We included three items for each video. Participants rated their approval to the statements "The movement of the ball on the right screen is fluid." and "The movement of the two balls is synchronous." on a 5-point Likert scale. Additionally, we included an adapted version of the ITU-R impairment scale [19]: Participants stated if they perceived a difference between both movements. The 5-point Likert scale ranged from "Imperceptible" to "Perceptible, but not annoying", "Slightly annoying", "Annoying" up to "Very annoying".

$N = 42$ people (19 female, 23 male) with a mean age of $M(SD) = 28.86(9.55)$ participated in the subjective evaluation. On average they reported playing video games $M(SD) = 6.7(10.24)$ hours a week with values ranging between 0 hours and 40 hours. 40 participants answered the employment question. 15 participants were students, 24 participants were employees, 1 participant was self-employed.

To analyze the data, we calculated a repeated-measures ANOVA for each item. For all three items, Mauchly's test indicated a violation of the assumption of sphericity (all $ps < .01$). Therefore, we report Greenhouse-Geisser-corrected tests for Fluidity ($\varepsilon = .74$), Synchrony ($\varepsilon = .66$), and Annoyment ($\varepsilon = .76$). All post-hoc tests were pairwise comparisons with Bonferroni adjustments. We used IBM SPSS Statistics 25 for the analysis of the quantitative data.

6.1 Results

Table 5 displays the descriptive statistics for the three items. Fig. 8 shows the means, standard errors, and significant differences. The ratings regarding the fluidity of the movement of the right ball differed significantly, $F(5.20, 213.05) = 43.74$, $p < .001$, partial $\eta^2 = .52$. Post-hoc tests showed that 2, 5, 10 and 25 avatars differed significantly from all higher numbers. 50 and 75 avatars differed from 100 avatars and higher ($p \leq .01$). No significant differences occurred between 2, 5, 10 and 25 avatars, between 50 and 75 avatars, and between 100 and 125 avatars. Participants' approval to the synchrony statement also differed significantly, $F(4.61, 188.89) = 101.15$, $p < .001$, partial $\eta^2 = .71$. Post-hoc tests revealed that 2, 5 and 10 avatars differed significantly from 50 and more avatars, 25 differed significantly from 75 and more, 50 and 75 avatars differed significantly from 100 and more, and 100 differed significantly from 125 (all $ps < .05$). Finally, the ratings on the ITU-R impairment scale (annoyment of the perceived difference) differed significantly, $F(5.31, 217.82) = 79.51, p < .001$, partial $\eta^2 = .66$. Post-hoc tests showed that the ratings differed significantly between the same numbers of avatars as for the synchrony ratings ($p \leq .01$). Overall, the subjective ratings were very much in line with the objective measures and did confirm the still acceptable limit of 25 avatars.

7 EP3 – Subjective Experience of Co-Location and Scalability

The aim of the final phase of the evaluation was three-fold: to get insights into (1) the subjective experiences of users immersed inside of an SVR with an increasing number of co-located embodied others, (2) the potential effects of different avatar appearances of the co-located others in such an environment, and (3) the impact of potential technical characteristics hampering the overall experience.

The user study followed a mixed-methods design. As the within-subjects factor, each participant experienced four conditions with a varying number of co-located avatars (2, 10, 25, 100) in randomized order. These numbers resulted from the first phases choosing 100 as a value certainly impacting the experience. As the between-groups factor, we manipulated the appearance of the other avatars. In the *Human* condition, all other avatars looked human (Fig. 3, right). In the *Mixed* condition, half of the avatars looked human, and the other half had an artificial appearance (Fig. 3, left). We assessed quantitative as well as qualitative data.

7.1 Procedure

Fig. 9 illustrates the experimental procedure. The first step introduced the participants to the procedure and the HMD and controller. Then they gave their informed consent to take part in the study and answered the pre-questionnaire. They put on the HMD and adjusted the head straps and lens distance according to their personal preferences.

Now participants experienced the first SVR scene consisting of 24 other avatars sitting around tables. Participants could inspect the surrounding for 20 seconds and then gave oral qualitative feedback on their impression of the scene without leaving the VR. Next, the experimenter showed an example question floating in front of the participant to explain how to interact with such in-vitro text questions in VR and to assure readability. The following experimental phase iterated through the four within-subject conditions, randomly varying the numbers of avatars. One abstract VR avatar sat at the same table as the participant throughout this phase. He moved the planets according to the recordings taken under the respective load condition. The participant answered questions in VR after each exposure. Fig. 10 shows screenshots of the scene (1) and a VR question afterward (2). In the end, participants removed the HMD and answered the post-questionnaire.

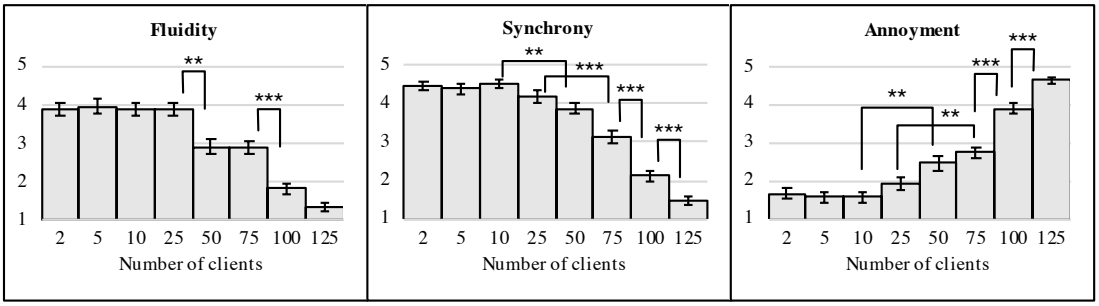

Fig. 8. Means and according standard errors for the items regarding the fluidity and synchrony of the movements and the annoyment. Low values mean low fluidity, synchrony, and annoyment. Significances are marked as follows: $^{*} < .05$, $^{**} < .01$, $^{***} < .001$.

The experiment simultaneously took place in three rooms with identical setups but different experimenters (2 male, 1 female). All experimenters followed a strict study protocol to ensure comparable results.

7.2 Measures

Participants filled in a pre- and post-questionnaire on a dedicated computer using the online questionnaire tool LimeSurvey and answered in-vitro questionnaires while immersed in the virtual environment.

1. Pre-Questionnaire: Participants answered the *Immersive Tendency Questionnaire* (ITQ) [50]. The ITQ consists of 18 items with 7-point Likert scales and values ranging from 1 to 7. The second part of the pre-questionnaire was the *Simulator Sickness Questionnaire* (SSQ) [22]. The questionnaire consists of 16 4-point scales ranging from 0 to 3.

2. Qualitative Feedback: To assess qualitative feedback, we asked the following questions:

- "How does it feel to be in this virtual environment?"
- "How do you feel about the presence of the others?"
- "Do you think you would interact with the others in the same way as you would in the real world?"

3. In-Vitro Questions: Presentation and answering of the in-vitro questions directly took place in the virtual environment after each condition. We measured the subjective presence of the participants with a single item as proposed in [7]. Participants answered the question "How present do you feel in the virtual environment right now?" on a rating scale ranging from 0 to 10. After that, participants stated their agreement on 7-point Likert scales ranging from 1 to 7 for the following five items:

- "The movement of the {ball / person at my table / other people in the room} was fluid."
- "The movement of the {person at my table / other people in the room} was natural."

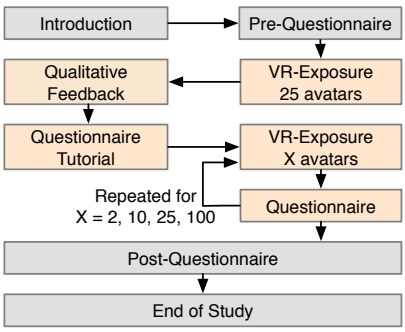

Fig. 9. General procedure of the experiment. Stages performed within VR are colored in orange.

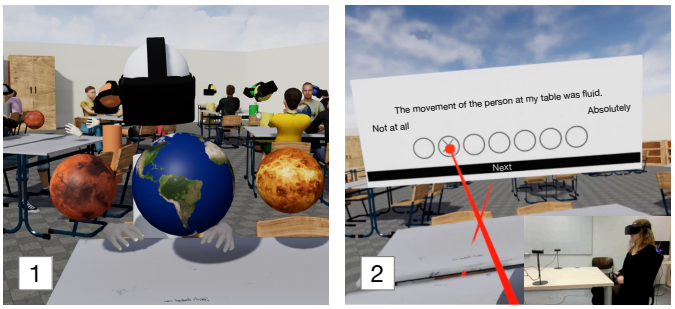

Fig. 10. The participants observed the movements of the planets for 30 seconds (1). They were asked to answer the questions directly in VR using a ray-cast pointing method (2). The participant selected an answer by pointing at a circular option field and pressing the forefinger button.

For the condition with only one additional avatar in the room, we excluded the items asking about the other people in the room. To obtain quantitative information about the experience of being in the virtual environment co-located with other avatars, we included the two subscales *co-presence* and *impression of interaction possibilities* of [34]. The co-presence subscale consists of three items, and the possible interactions subscale consists of four items. All scales are 7-point Likert scales ranging from 1 to 7.

4. Post-Questionnaire: To gather information about possible uncanny valley effects we included the four subscales *Attractiveness*, *Humanness*, and *Eeriness* [16, 17]. We divided Eeriness into the subscales *eerie* and *spine-tingling* [16]. The last stimulus-related question was another open question: "How did you feel about the different numbers of avatars in the room?"

Participants answered the SSQ [22] a second time to assure that our application did not induce any unwanted effects regarding simulator sickness. Each participant answered the following demographic items: age, gender, occupation, highest educational achievement, debility of sight, years of experience with the location's language, and experience in playing video games and with VR environments.

7.3 Participants

$N = 47$ people participated in the study. We excluded two data sets from the analysis due to technical problems (1 disconnected controller, 1 false position calibration) during the experiment. The remaining sample consisted of $N = 45$ people (71.7 % female) with a mean age of $M(SD) = 20.96(1.75)$. All participants gave written informed consent and got course credits for their participation. Assignment of participants to one of the two conditions Human ($n = 23$, 56.5% female) or Mixed ($n = 22$, 86.4% female) was randomized. In both groups, most people reported playing video games less than 1 hour a day (Human: 14, Mixed: 18 people). In the Mixed condition, four people stated that they had never experienced Virtual Reality with a head-mounted

Table 6. Descriptive statistics for the Pre- and Post-Questionnaire Scales. ITQ-scales range from 1 to 7. SSQ scores are total scores calculated according to [22]. For the uncanny valley subscales (value range -3 to +3) low scores mean low humanness, low eeriness, and low attractiveness.

	Human	**Mixed**
Questionnaire	$M(SD)$	$M(SD)$
ITQ	4.35(.67)	4.66(.49)
SSQ Pre	12.68(14.56)	14.96(15.31)
SSQ Post	12.52(14.82)	13.60(15.68)
Uncanny Valley	**Human**	**Mixed**
Attractiveness	.52(.91)	1.07(.71)
Humanness	-.39(1.37)	.30(.98)
Overall Eeriness	-1.07(.65)	-.58(.68)
Eerie	-.88(.94)	-.44(.65)
Spine-Tingling	-1.22(.60)	-.69(.89)

display before. In both groups, about 50 % stated that they have 1 to 5 hours of VR experience. In the Human condition, more people (7) stated to have more than 10 hours of VR experience than in the Mixed Condition (3 people). The two groups did not differ significantly regarding their Immersive Tendency ($t(43) = -1.75, p = .09$) or their Simulator Sickness Ratings before ($t(43) = -.51, p = .61$) or after ($t(43) = -.24, p = .81$) the experiment. Overall, simulator sickness did not change significantly throughout the experiment, $t(44) = .53, p = .60$. Table 6 displays the descriptive statistics for the ITQ and SSQ.

7.4 Results

We used IBM SPSS Statistics 25 for the analysis of the quantitative data. Table 7 reports descriptive statistics for all dependent variables assessed inside of the virtual environment.

7.4.1 In-Vitro Measurements

We calculated mixed design ANOVAs for all measurements assessed inside of the virtual environment. Mauchly's test indicated a violation of the assumption of sphericity for the fluidity of the ball ($\varepsilon = .61$), the fluidity ($\varepsilon = .53$) and naturalness ($\varepsilon = .64$) of the person at the participant's table, as well as for the perceived possibility of interaction ($\varepsilon = .71$), with all $ps < .001$. Therefore, we report Greenhouse-Geisser-corrected tests for these variables. For the fluidity and naturalness of the crowd as well as for presence and co-presence Mauchly's test indicated that sphericity could be assumed.

The analysis reveals a significant main effect of the number of clients on the perceived fluidity of the ball ($F(1.82, 78.23) = 33.30$, partial $\eta^2 = .44$) and the perceived fluidity ($F(1.60, 68.94) = 42.61$, partial $\eta^2 = .50$) and naturalness ($F(1.92, 82.48) = 27.83$, partial $\eta^2 = .39$) of the person at the participant's table, all $ps < .001$. Post-hoc tests for these main effects show that the condition with 100 clients differs significantly from all other conditions (all $ps < .001$). Participants rated this condition the least fluid (ball and person at the table) and natural. We found no significant main effects of the number of avatars on the perceived fluidity or naturalness of the crowd. In all tests regarding the fluidity and naturalness, we found no significant main effect of the avatar appearance (Human vs. Mixed).

The main effect of the number of avatars on the presence rating was not significant, $F(3, 129) < 1$, $p > .05$, partial $\eta^2 = .02$. Presence ratings were similar across conditions with means ranging between $M = 6.22$ and $M = 7.09$. Presence was rated highest in the condition with 25 avatars. There was a significant main effect of the number of avatars on the perceived co-presence, $F(3, 129) = 22.84$, $p < .001$, partial $\eta^2 = .35$. In the condition with only two avatars (the participant plus the person at the table) co-presence was rated lowest, differing significantly (all $ps < .001$) from all other conditions. The co-presence ratings for 10, 25, and 100 avatars do not differ significantly and are similarly high. We found a significant interaction between the number and the appearance of the clients regarding co-presence, $F(3, 129) = 2.85$, $p < .05$, partial $\eta^2 = .06$. The difference between the condition with two avatars and the other three conditions was smaller in the group that saw the human avatar crowd. Compared to this group, the group with the mixed avatar crowd gave a lower co-presence rating for the condition with only one additional avatar and higher ratings for the other three numbers of avatars. We also found a significant main effect of the numbers of clients on the perceived possibility of interactions, $F(2.13, 91.57) = 6.19$, $p = .002$, partial $\eta^2 = .13$. Post-hoc tests show that the condition with 25 clients differs significantly from the condition with two clients ($p = .002$) and the conditions with 100 clients ($p = .046$). The 25 clients condition shows the highest rating regarding the perceived possibility of interaction. For presence, co-presence, and the perceived possibility of interactions nearly all ratings were higher in the group that saw the mixed avatar appearances. However, we found no significant main effect of the avatar appearance (Human vs. Mixed).

7.4.2 Uncanny Valley

We compared the ratings for attractiveness, humanness, and eerieness between both groups calculating independent t-tests. Levene-tests for all subscales were non-significant. The groups did not differ significantly regarding the perceived humanness of the avatars, $t(43) = -1.95, p = .06, r = .29$. Nevertheless, it is noteworthy that the perception of the humanness of the human-looking avatar crowd was lower than of the mixed avatar crowd.

The avatar crowd with mixed appearances appeared as significantly more attractive than the human avatar crowd $t(43) = -2.23, p < .05$, $r = .32$. The two groups differed significantly regarding the perceived eeriness. The human looking avatar crowd appeared to be less eerie than the mixed avatar crowd, $t(43) = -2.47, p < .05$, $r = .35$. As proposed in [16], we split eeriness into its two subscales *eerie* and *spine-tingling*. As a result, we found no significant difference regarding the eerie subscale, $t(43) = -1.81, p > .05$, $r = .27$, but the human looking avatar-crowd was significantly less spine-tingling than the mixed avatar crowd, $t(43) = -2.34, p < .05, r = .34$.

7.4.3 Qualitative Feedback

Qualitative feedback was mixed. Some users were surprised by how real the whole scenario felt and how real the avatars appeared to be. Other users made exactly opposite remarks. Comments about the realness of the avatars were usually restricted to the humanlike avatars. The abstract avatars were described as "things" or robots. Participants almost consistently denied them human status. Many participants commented on the movements of the avatars. Some noticed the repetition of movements. Many stated that the movements seem unnatural.

Some participants reported to feel alone or as being excluded: The avatars shared their tables with others while the participant was the silent observer in the middle. Some said the other avatars ignored them while others interpreted the avatars as staring at them. When getting accustomed to the situations, they reported that more avatars decreased the feeling of loneliness. The situation was reported to be overwhelming when too many avatars were present. Some participants would have liked to interact with the avatars if it were possible because they looked real. Others rejected a potential interaction because they did not experience the avatars as real enough. In both cases, the stated decision factor was how humanlike the avatar is perceived.

8 Discussion and Conclusion

We defined specific requirements for SVRs to evaluate how scalability would affect overall subjective and objective system performance. We developed a software system with consumer soft- and hardware to identify the current state-of-the-art with such an approach. The system design includes all the functional requirements initially defined by R1 to R6. The system supports scalability in the number of distributed co-located avatars, the sensory coverage, as well as in the variable avatar appearance up to photorealistic avatars created by photogrammetry.

Benchmarking confirmed the non-functional performance requirements using objective performance characteristics. We demonstrated how the latter coherently matches the user experience measured by

Table 7. Descriptive statistics for all dependent variables assessed inside the virtual environment (in-vitro). Item scales for fluidity, naturalness, co-presence and interaction range from 1 to 7. The mid-immersion presence item ranges from 0 to 10.

In-Vitro Questions	Condition	2 clients $M(SD)$	10 clients $M(SD)$	25 clients $M(SD)$	100 clients $M(SD)$
Fluid Ball	Human	5.48(1.34)	5.87(1.06)	5.57(1.34)	3.30(1.82)
	Mixed	5.36(1.29)	5.18(1.33)	5.45(1.34)	3.68(2.10)
Fluid Person Table	Human	5.48(1.34)	5.74(.96)	5.61(1.23)	3.22(1.54)
	Mixed	5.82(1.05)	5.59(.85)	5.59(1.01)	4.00(2.00)
Fluid Others	Human	-	4.04(1.55)	4.13(1.58)	4.43(1.44)
	Mixed	-	4.77(1.15)	4.64(1.29)	4.82(1.18)
Natural Person Table	Human	5.13(1.29)	5.48(.90)	5.26(1.21)	3.96(1.55)
	Mixed	5.14(1.25)	5.27 (1.03)	5.18(1.47)	3.68(1.46)
Natural Others	Human	-	3.52(1.56)	3.70(1.58)	3.35(1.61)
	Mixed	-	4.73(1.03)	4.36(1.36)	4.55(1.22)
Presence	Human	6.39(2.13)	6.39(2.23)	6.57(2.79)	6.22(2.73)
	Mixed	6.73(1.32)	6.82(1.30)	7.09(1.77)	6.64(1.99)
Co-Presence	Human	3.81(1.87)	4.90(1.52)	4.87(1.66)	4.88(1.61)
	Mixed	3.26(1.63)	5.46(.71)	5.46(.99)	5.58(1.09)
Interaction	Human	2.52(1.38)	2.74(1.18)	2.99(1.31)	2.58(1.20)
	Mixed	2.55(1.40)	3.32(1.26)	3.52(1.29)	3.16(1.19)

subjective user reports on perceived system characteristics in the non-immersive as well as in the immersive setup. Here, it only affected the perception of the modified movements of the ball and the interaction partner as assumed. Our evaluations also confirm the theoretical estimates of the bandwidth requirements for the client-server system, and the chosen interconnect and fidelity B3 (324.0 KB/s) with a maximum of 25 concurrently co-located distributed avatars, leaving enough bandwidth to distribute whole bodies as specified by B4 and to avoid client-side IK, if required.

We investigated the experience of users immersed inside an embodied SVR with a variable number of participants and appearances of avatars. The condition with 25 avatars significantly resulted in the highest perceived possibility of interaction and had the highest presence ratings. Co-presence was significantly lower for the two avatar condition, and there was a significant interaction between number and appearance of the crowds. Here, potential inconsistencies and incoherences with participants' expectations may cause them to more intensely focus on the surrounding avatars, which would be in line with the significantly higher attractiveness of the mixed avatar crowd.

In general, the vivid surrounding with active companions not hampered by any technical limitation (as emerging here for 25+ avatars) seems to imply a dynamic and stimulating environment despite the used canned animations. Also in line with the reported significant results, presence, the possibility of interaction, and co-presence were consistently higher for the mixed crowd.

All results confirm the positive effects of co-located social companions as well as detrimental effects of suboptimal system performance (here illustrated for 25+ avatars) on the quality of experience of virtual worlds. Finally and notably, the human avatars were rated less human but also significantly less eerie than the mixed crowd. We explain this discrepancy by the incoherence between static appearance and behavior appearance. Overall, these results also inspire an interesting design finding: If we want to manipulate users' interest into a given SVR we can do so by providing mixed avatar appearances, but we have to consider that we are also increasing an inherent eeriness, which might or might not be advisable for a given application context.

8.1 Future Work

Future work will test performance and user experience for extended avatar fidelities and appearances including speech interaction and will experiment with various optimization schemes. These evaluations will be followed by studying application-specific effectiveness for training and learning inside an SVR.

Acknowledgments

This work was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF project *ViLeArn*).

References

[1] J. Achenbach, T. Waltemate, M. E. Latoschik, and M. Botsch. Fast generation of realistic virtual humans. In *23rd ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 12:1–12:10, 2017.

[2] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments*, 15(4):359–372, 2006.

[3] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):616–625, 2013.

[4] G. Bente, S. Rüggenberg, N. C. Krämer, and F. Eschenburg. Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human communication research*, 34(2):287–318, 2008.

[5] C. Blanchard, S. Burgess, Y. Harvill, J. Lanier, A. Lasko, M. Oberman, and M. Teitel. Reality built for two: a virtual reality tool. In *ACM SIGGRAPH Computer Graphics*, volume 24, pages 35–36. ACM, 1990.

[6] A. Bönsch, S. Radke, H. Overath, L. M. Asché, J. Wendt, T. Vierjahn, U. Habel, and T. W. Kuhlen. Social VR: How personal space is affected by virtual agents' emotions. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 199–206, 2018.

[7] S. Bouchard, J. St-Jacques, G. Robillard, and P. Renaud. Anxiety increases the feeling of presence in virtual reality. *Presence: Teleoperators and Virtual Environments*, 17(4):376–391, 2008.

[8] C.-M. Chang, C.-H. Hsu, C.-F. Hsu, and K.-T. Chen. Performance measurements of virtual reality systems: Quantifying the timing and positioning accuracy. In *Proceedings of the 24th ACM international conference on Multimedia*, MM '16, pages 655–659, New York, NY, USA, 2016. ACM.

[9] M. Chollet and S. Scherer. Perception of virtual audiences. *IEEE Computer Graphics and Applications*, 37(4):50–59, 2017.

[10] M. Dou, H. Fuchs, and J.-M. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 99–106. IEEE, 2013.

[11] P. Ekman and W. V. Friesen. *Manual for the facial action coding system*, 1978.

[12] S. Friston and A. Steed. Measuring latency in virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 20(4):616–625, 2014.

[13] M. Gross, M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. V. Moere, and O. Staadt. Blue-c: A spatially immersive display and 3d video portal for telepresence. *ACM Transactions on Graphics (TOG)*, 22(3):819–827, 2003.

[14] D. He, F. Liu, D. Pape, G. Dawe, and D. Sandin. Video-based measurement of system latency. In *International Immersive Projection Technology Workshop*, 2000.

[15] P. Heidicker, E. Langbehn, and F. Steinicke. Influence of avatar appearance

on presence in social VR. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 233–234, 2017.

[16] C.-C. Ho and K. F. MacDorman. Measuring the uncanny valley effect. *International Journal of Social Robotics*, 9(1):129–139, 2017.

[17] C.-C. Ho, K. F. MacDorman, and Z. A. D. Pramono. Human emotion and the uncanny valley: a GLM, MDS, and Isomap analysis of robot video ratings. In *3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 169–176. IEEE, 2008.

[18] W. A. IJsselsteijn, Y. A. W. de Kort, and A. Haans. Is this my hand I see before me? The rubber hand illusion in reality, virtual reality, and mixed reality. *Presence: Teleoperators and Virtual Environments*, 15(4):455–464, 2006.

[19] ITU. Methodology for the subjective assessment of the quality of television pictures, recommendation ITU-R BT. 500-11. Technical report, ITU Telecom. Standardization Sector of ITU, 2002.

[20] D. Jeffers. Is there a second life in your future? In *Proceedings of the 36th Annual ACM SIGUCCS Fall Conference: Moving Mountains, Blazing Trails*, pages 187–190, New York, NY, USA, 2008. ACM.

[21] V. Kasapakis, E. Dzardanova, and C. Paschalidis. Conceptual and technical aspects of full-body motion support in virtual and mixed reality. In L. T. De Paolis and P. Bourdot, editors, *International Conference on Augmented Reality, Virtual Reality, and Computer Graphics*, pages 668–682. Springer International Publishing, 2018.

[22] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3):203–220, 1993.

[23] M. E. Latoschik, J.-L. Lugrin, M. Habel, D. Roth, C. Seufert, and S. Grafe. Breaking bad behavior: Immersive training of class room management. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology (VRST)*, pages 317–318, New York, NY, USA, 2016. ACM.

[24] M. E. Latoschik, J.-L. Lugrin, and D. Roth. FakeMi: A Fake Mirror System for Avatar Embodiment Studies. In *Proceeding of the 22nd ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 73–76, 2016.

[25] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch. The effect of avatar realism in immersive social virtual realities. In *23rd ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 39:1–39:10, 2017.

[26] Y. Liu, S. Beck, R. Wang, J. Li, H. Xu, S. Yao, X. Tong, and B. Froehlich. Hybrid lossless-lossy compression for real-time depth-sensor streams in 3D telepresence applications. In Y.-S. Ho, J. Sang, Y. M. Ro, J. Kim, and F. Wu, editors, *Advances in Multimedia Information Processing – PCM 2015*, pages 442–452, Cham, 2015. Springer International Publishing.

[27] J.-L. Lugrin, F. Charles, M. Habel, H. Dudaczy, S. Oberdörfer, J. Matthews, J. Porteous, A. Wittmann, C. Seufert, S. Grafe, and M. E. Latoschik. Benchmark framework for virtual students' behaviours. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2236–2238, 2018.

[28] J.-L. Lugrin, M. Ertl, P. Krop, R. Klüpfel, S. Stierstorfer, B. Weisz, M. Rück, J. Schmitt, N. Schmidt, and M. E. Latoschik. Any "body" there? Avatar visibility effects in a virtual reality game. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 17–24. IEEE, 2018.

[29] J.-L. Lugrin, M. E. Latoschik, M. Habel, D. Roth, C. Seufert, and S. Grafe. Breaking bad behaviours: A new tool for learning classroom management using virtual reality. *Frontiers in ICT*, 3:26, 2016.

[30] J.-L. Lugrin, J. Latt, and M. E. Latoschik. Anthropomorphism and illusion of virtual body ownership. In *Proceedings of the 25th International Conference on Artificial Reality and Telexistence and 20th Eurographics Symposium on Virtual Environments*, pages 1–8. Eurographics Association, 2015.

[31] A. Maselli and M. Slater. The building blocks of the full body ownership illusion. *Frontiers in human neuroscience*, 7:83, 2013.

[32] J. McVeigh-Schultz, E. Márquez Segura, N. Merrill, and K. Isbister. What's it mean to "Be Social" in VR?: Mapping the social VR design ecology. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems (DIS)*, pages 289–294. ACM, 2018.

[33] D.-P. Pertaub, M. Slater, and C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators & Virtual Environments*, 11(1):68–78, 2002.

[34] S. Poeschl and N. Doering. Measuring co-presence and social presence in virtual environments–psychometric construction of a German scale for a fear of public speaking scenario. *Annual Review of Cybertherapy and Telemedicine*, pages 58–63, 2015.

[35] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 179–188. ACM, 1998.

[36] S. Rehfeld, M. E. Latoschik, and H. Tramberend. Estimating latency and concurrency of asynchronous real-time interactive systems using model checking. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 57–66. IEEE, 2016.

[37] D. Roberts, R. Wolff, J. Rae, A. Steed, R. Aspin, M. McIntyre, A. Pena, O. Oyekoya, and W. Steptoe. Communicating eye-gaze across a distance: Comparing an eye-gaze enabled immersive collaborative virtual environment, aligned video conferencing, and being together. In *IEEE Virtual Reality Conference*, pages 135–142, 2009.

[38] D. Roth, C. Kleinbeck, T. Feigl, C. Mutschler, and M. E. Latoschik. Beyond Replication: Augmenting Social Behaviors in Multi-User Social Virtual Realities. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 215–222, 2018.

[39] D. Roth, K. Waldow, F. Stetter, G. Bente, M. E. Latoschik, and A. Fuhrmann. SIAMC – A socially immersive avatar mediated communication platform. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology (VRST)*, pages 357–358. ACM, 2016.

[40] R. Schroeder. *Being There Together: Social interaction in shared virtual environments*. Oxford University Press, 2010.

[41] M. Slater. Grand challenges in virtual environments. *Frontiers in Robotics and AI*, 1:3, 2014.

[42] M. Slater, D. Perez-Marcos, H. Ehrsson, and M. V. Sánchez-Vives. Towards a digital body: the virtual arm illusion. *Frontiers in Human Neuroscience*, 2(6), 2008.

[43] M. Slater, D.-P. Pertaub, and A. Steed. Public speaking in virtual reality: Facing an audience of avatars. *IEEE Computer Graphics and Applications*, 19(2):6–9, 1999.

[44] H. J. Smith and M. Neff. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 289:1–289:12. ACM, 2018.

[45] B. Spanlang, J.-M. Normand, D. Borland, K. Kilteni, E. Giannopoulos, A. Pomés, M. González-Franco, D. Perez-Marcos, J. Arroyo-Palacios, X. N. Muncunill, and M. Slater. How to build an embodiment lab: Achieving body representation illusions in virtual reality. *Frontiers in Robotics and AI*, 1:9, 2014.

[46] A. Steed. A simple method for estimating the latency of interactive, real-time graphics simulations. In *Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 123–129, New York, NY, USA, 2008. ACM.

[47] A. Steed and R. Schroeder. Collaboration in Immersive and Non-immersive Virtual Environments. In *Immersed in Media*, pages 263–282. Springer, 2015.

[48] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1643–1652, 2018.

[49] T. Waltemate, F. Hülsmann, T. Pfeiffer, S. Kopp, and M. Botsch. Realizing a low-latency virtual reality environment for motor learning. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 139–147. ACM, 2015.

[50] B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and virtual environments*, 7(3):225–240, 1998.

[51] N. Yee and J. Bailenson. The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, 33(3):271–290, 2007.