A Toolkit for Creating Intelligent Virtual Humans in Extended Reality

Fariba Mostajeran[1a*], Ke Li[1b*], Sebastian Rings[1], Lucie Kruse[1], Erik Wolf[1], Susanne Schmidt[1], Michael Arz[1], Joan Llobera[2], Pierre Nagorny[2], Caecilia Charbonnier[2], Hannes Fassold[3], Xenxo Alvarez[4], André Tavares[4], Nuno Santos[4], João Orvalho[4], Sergi Fernández[5], Frank Steinicke[1]

[1]University of Hamburg, [2]ARTANIM Foundation, [3]JOANNEUM RESEARCH, [4]Didimo, [5]i2CAT Foundation

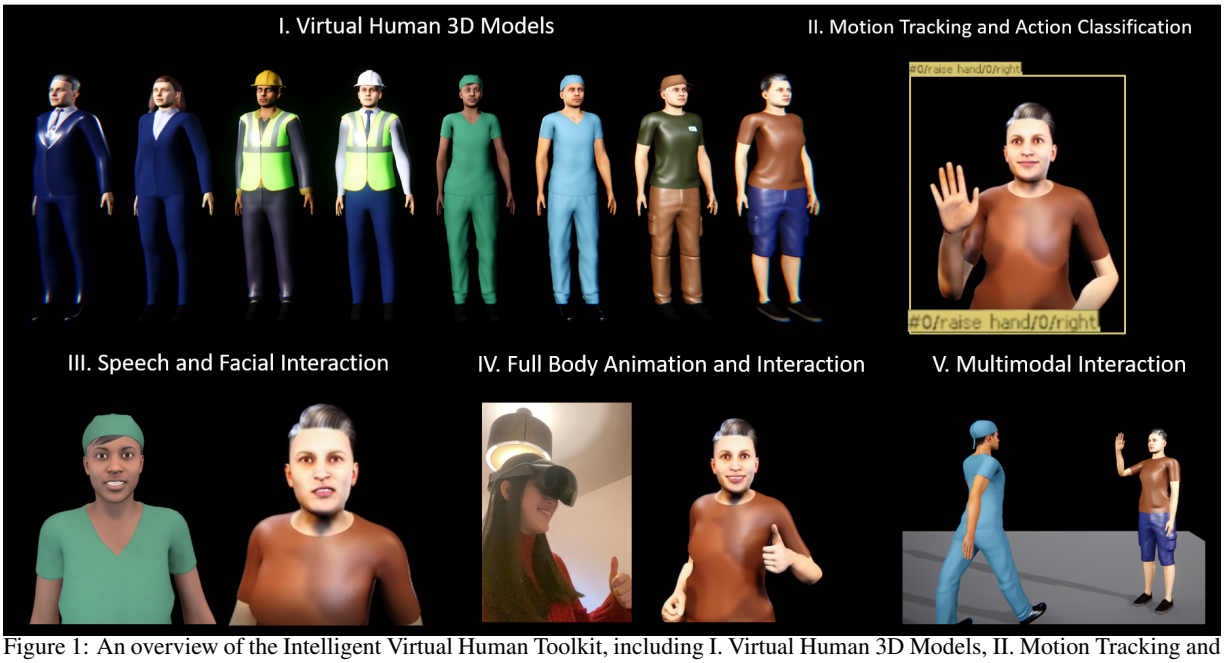

Figure 1: An overview of the Intelligent Virtual Human Toolkit, including I. Virtual Human 3D Models, II. Motion Tracking and Action Classification; III. Speech and Facial Interaction, IV. Full Body Animation and Interaction, and V. Multimodal Interaction modules.

Abstract

This paper introduces an initial implementation and a preliminary evaluation of a toolkit for providing Intelligent Virtual Humans in Extended Reality (XR) environments. These virtual humans may serve as avatars for users, known as Smart Avatars, or artificial agents, known as Intelligent Virtual Agents. The presented toolkit consists of five modules that contribute to producing realistic human-to-human and human-to-agent interactions in XR. The primary focus lies in enabling natural, multimodal communication and interaction through a range of expressive capabilities, including speech, facial expressions, gaze tracking, and full-body animations. By enhancing the realism and responsiveness of virtual humans, this paper strives to facilitate their use in a variety of metaverse applications.

Index Terms: Extended Reality, Intelligent Virtual Agents, Smart Avatars

*These authors contributed equally to the work.
[b]e-mail: fariba.mostajeran.gourtani@uni-hamburg.de
[a] e-mail: ke.li@uni-hamburg.de

1 Introduction

An essential aspect of the Metaverse is to provide users with a representation of themselves and 'others' in Extended Reality (XR) environments [36]. The term 'others' refers to both humans, represented as (smart) avatars, as well as intelligent virtual agents (IVAs). These intelligent virtual humans play a crucial role in various XR applications, including cultural heritage [15], training [53], health care [21, 2], and remote collaborations [50].

Despite their significant potential, creating intelligent virtual humans for XR applications faces various challenges. Firstly, to give users a high sense of presence and immersion in an XR experience, the 3D human models need to accurately represent the visual and physical appearances of the users so that they can be reliably used as their 3D virtual avatars [47]. The process of generating these 3D characters must also be fast and cost-effective, enabling users to create personalized virtual representations easily [10]. Secondly, to enable engaging human-agent interactions and collaborations, IVAs need to perform human-like behaviors that are both plausible and natural. For example, the IVAs can detect and understand human users' actions and create suitable multimodal responses to users' multimodal inputs via natural speech, facial, and full-body interactions [6]. Moreover, virtual humans not only need to behave in a socially-compatible manner but their motions and interactions in the virtual world need to follow the physical laws like that of the real world, thereby, ensuring a high level of embodiment and presence during the XR experience [49].

This paper introduces our ongoing research to design and imple-

ment intelligent virtual humans targeting cost-efficient approaches. Our goal is to create a comprehensive toolkit for both smart avatars (SAs) and intelligent virtual agents (IVAs), supporting: (I) a customizable 3D human model creation pipeline, (II) low-latency motion tracking and action classification through machine learning (ML) integration, (III) natural speech and facial expression synthesis, and (IV) realistic, physics-based full-body animations and interactions (see Figure 1). Our goal is to facilitate natural and efficient multimodal interactions between SAs and IVAs, enabling their application across a wide range of metaverse use cases. We provide an initial implementation of the toolkit, along with a preliminary evaluation involving four XR developers. We also discuss the current limitations and challenges encountered during development and outline a roadmap for the toolkit's continuous improvement and integration in future releases.

2 Related Work

Intelligent Virtual Agents (IVAs) are autonomous characters designed to interact naturally with humans in virtual environments. Recent advances in artificial intelligence have significantly enhanced their capabilities across several dimensions. IVAs typically integrate natural language processing for conversation, computer vision for environmental awareness, and generative models for behavior generation. Notable developments include improved emotional intelligence through facial expression recognition and generation [23], context-aware dialogue systems [25], and sophisticated decision-making architectures [28] . Their effectiveness depends heavily on the seamless integration of verbal and non-verbal communication channels, including gaze, gestures, and full-body movements.

Previous studies have shown that humans maintain social rules in the presence of more human-like virtual agents [32]. Bailenson et al. [3], for example, showed that people in virtual environments keep a larger distance to virtual agents than to virtual objects. They also associate more human-like characteristics such as being alive, calm, intelligent, and friendly to virtual agents in XR [33, 35]. Moreover, research has shown that IVAs can be used to elicit human emotions [39] such as psycho-social anxiety [31] or facilitate their cognitive [22] or physical task performances [34].

On the other hand, avatars are referred to as virtual characters that are controlled by real users and are used for self-representation in virtual worlds [12]. Previous studies have shown that the sense of presence and embodiment can be improved when users are provided with avatars [44]. Avatars can facilitate generating the body-ownership illusion which arises when users have a sense of ownership over the virtual body that they have received in the virtual world, despite the certain knowledge that the virtual body is not their real body [29]. When using a humanoid avatar, users typically receive an upper body representation which can be controlled with limited input including a head-mounted display (HMD) and hand controllers. However, recent research has proposed using Smart Avatars (SA) [13] which can perform complex movements and express natural behavior despite having limited system input. For instance, users with SAs can have continuous full-body human representations for noncontinuous locomotion in XR. In addition, if the users teleport, their SAs would imitate their assigned user's real-world movements and autonomously navigate to their user when the distance between them exceeds a certain threshold. Thus, the observers could observe a natural human walk for that user instead of instant jumps caused by the teleportation.

A number of methods have been employed to create realistic humanoid 3D models that can be used as both agents and avatars. This may include complex technical setups, such as multi-view camera domes [7], or AI-based approaches, such as generative neural network architectures and diffusion models (e.g., autoencoders [27]), and Neural Rendering. The combination of these methods can potentially overcome the limitations of current solutions (Unity ZIVA[1], Unreal MetaHuman[2] or SoulMachines[3]), such as disturbing gaps in the perception process leading to uncanny valley effects [30], particularly noticeable for facial animations given that the human neural system is extremely sensitive for processing faces [20].

Regarding interactive character animation, recent years have shown considerable progress in the use of machine learning techniques for both kinematic (see, for example, [51, 42]) and physics-based controllers (see, for example, [38, 4, 16, 24, 52]). These have made these techniques more amenable to the creation of IVAs for VR experiences. It also allows exploring the perceived quality of character animation in VR experiences [8], and introduces a different way to investigate open questions in motor neuroscience [26].

There are however still numerous issues to address in the field. Current techniques can generate physically plausible movements [38, 37], but achieving the nuanced expressivity and stylistic variations typical of human motion remains difficult [37, 48]. The mastering of body language is still a challenge, maintaining a consistent character style across different behaviors [1, 41]. To enable interactions with embodied users, real-time performance is mandatory to react with the lowest latency [45]. Both users and environmental interactions present exciting opportunities for future research in combining physics-based approaches with data-driven methods to create more sophisticated and believable virtual characters [43, 4, 16].

3 Intelligent Virtual Human Toolkit

In this work, we present our Intelligent Virtual Human (IVH) toolkit whose goals are i) to provide real-time photorealistic humanoid 3D models based on cost-efficient technology, which can represent users (avatars) or agents (IVA), ii) to support natural and multimodal communication and interaction via speech, facial expressions, gaze, and full-body animations, and iii) to move from simple human-human communication to hybrid forms of interaction including multiple real and artificial users represented by smart avatars and IVAs. This toolkit is developed for the Unity3D game engine and comprises five modules that are explained in this section.

I. Virtual Human 3D Models

The first module provides a set of virtual humanoid 3D models, diverse in terms of represented gender, age, and ethnicity. They are intended to be used out of the box for multiple use cases including professional collaborations, manufacturing training, health care, and cultural heritage (see Figure 1-I). In addition, this module will deliver an efficient user-driven offline pipeline for generating fully rigged, skinned, and animatable humanoid 3D models with high visual fidelity [9]. It will use simple (mobile) camera setups to capture RGB-depth photo of the user's head to serve as an input to a convolutional neural network that estimates the weights of a morphable model to produce an initial head shape that is further adjusted through landmark-guided deformation [10]. When creating the head model, the input data will be normalized in terms of specularity, shadows, and image artifacts, and relighting methods will be implemented to ensure that the generated virtual humans can be used in different virtual environments adapting to various environmental conditions.

II. Motion Tracking and Action Classification

This module is responsible for detecting and classifying the actions of virtual humans in the scene in realtime. Unfortunately, the machine learning capabilities available within Unity are severely limited. For example, the *Barracuda* Unity package for neural networks does not support modern transformer-based architectures.

[1] https://unity.com/blog/news/update-about-ziva
[2] https://www.unrealengine.com/en-US/metahuman
[3] https://www.soulmachines.com/

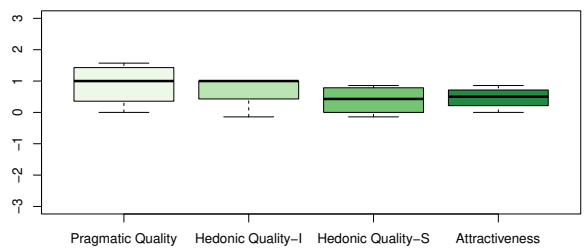

Figure 2: Mean values relating to the AttrakDiff sub-scales, including Pragmatic Quality (PQ), the Hedonic Quality concerning Identity (HQ-I), the Hedonic Quality concerning Stimulation (HQ-S), and the general Attractiveness (ATT).

We therefore developed a novel framework that relies on two main components: 1) A *Visual Analyzer* Python application that does the actual realtime action classification and sends the classification results to Unity, and 2) a Unity package that receives the result of the action classification via REST API and renders a video live stream from a certain virtual camera viewpoint. In that way, we can use the powerful Python packages for deep learning (like Pytorch, etc.) as well as helpful Python modules for parallel processing like the *Multiprocessing* module. The Visual Analyzer application does the processing of the input video stream received from Unity in multiple steps. First, all virtual humans are detected with an object detector (we employ *Scaled-YoloV4*) and tracked with an optical-flow-based method. In parallel, for all detected humans their 2D pose (skeleton) is calculated with the *RTMPose* pose estimation algorithm. The actions of one human are now calculated from the trajectory of its 2D poses within approximately the last second. Currently, as shown in Figure 1 II, the module can already detect the action when a virtual human raises a hand in realtime. For detecting this action, we use a simple heuristic: we check whether the elbow is located higher than the torso and whether the upper arm is approximately vertically oriented. The algorithm can detect this action with a delay of less than one second. In the future, we will integrate deep-learning-based action classification algorithms which are also using the 2D skeletons as input.

III. Speech and Facial Interaction

This module focuses on enhancing the communication capabilities of the IVHs through speech and facial interaction utilizing the 3D humanoid characters developed in the first module. This module aims to achieve two main objectives using the models created in Module I: the creation of SAs, which embody users in XR environments, and the development of IVAs, which are AI-powered computer-generated characters. These avatars and agents enhance interactivity and engagement in XR applications, utilizing advanced AI technologies for natural communication.

As shown in Figure 1 III, we have developed methods for natural communication between humans and SAs as well as IVAs based on processing spoken language and facial expressions. We have integrated AI-based services to enable speech-based interaction between users and IVAs. In particular, we have employed speech-to-text (currently Google API STT) and text-to-speech synthesis (currently OpenAI TTS and Azure TTS) technologies and large language models (currently through OpenAI's chat creation endpoint using models such as GPT-4o, GPT-4o-mini, or GPT3.5-turbo) to provide conversational capabilities for the IVAs. This allows for contextually relevant and coherent dialogues between users and IVAs. With the training of a custom or personal neural voice through Azure, it is possible to further increase the realism of digital twins, also including the voice of the user. The SAs of the users have also facial expression and upper-body tracking (i.e., head and hands) leveraging the latest development of the *Meta Movement SDK*[4]. Both types of IVHs benefit from Oculus' lipsync technology[5] which maps users' or IVAs' voices to the movements of their virtual lip representations in XR. Finally, the IVAs are capable of expressing six basic emotions (happiness, sadness, anger, disgust, fear, and surprise) in different intensities and durations according to their conversation with the user. This is done by changing their corresponding facial blendshapes based on the action units of the facial action coding system (FACS) [11].

IV. Full Body Animation and Interaction

To animate IVAs, we are exploring the extent to which emerging physics-based motor control techniques can improve the body movements of interactive virtual reality (VR) characters. The technical challenge is to bring physics-based character animation techniques – typically used in robotics simulations – to control the behavior of the IVAs in VR production environments. For this purpose, we have improved a set of reinforcement learning environments implemented in Unity, using ML-agents for reinforcement learning, which are available online [6] as an open source project called *Modular Agents*. We have also adapted these environments to work within the toolkit for specific scenarios (see section Multimodal Interaction). Our overall goal is to provide tools to train physics-based character controllers and make them accessible to a broader community, beyond academic researchers. To do this we are exploring training these controllers both within the VR development environment and externally, on a separate physics simulation engine. We aim to optimize performance to support real-time VR applications. The trained controllers are exported as neural network policies that can run efficiently within Unity, following established practices for real-time neural network deployment [19, 14], maintaining physical accuracy while meeting the strict performance requirements of XR applications.

V. Multimodal Interaction

This module is dedicated to integrating the outcomes from the previously mentioned modules to create IVHs capable of realistic multimodal communication. As shown in Figure 1 V, in an exemplary scenario, the user has a smart avatar, sees a standing IVA in the distance, and greets the agent while raising their hand (comprising I and III modules). By hearing this and detecting the raised hand (module II), the agent approaches the user, respects the social distance (module IV), and greets the user back. The conversation can continue (module III). If the user changes the social distance and comes closer to the agent, the agent shall move back and resume the socially acceptable distance.

4 Preliminary Evaluation

We presented and provided a version of our IVH toolkit which contained modules I (with only one virtual humanoid model) and III to the Computer Science students of the Department of Computer Science at the University of Hamburg. They used this version to develop their Unity projects about IVHs for their Master's project. Four groups of students were working on different topics, all including IVAs. Their task was to develop a research study on I) natural interruption techniques, II) referencing scene objects in a conversation with an IVA, III) non-verbal communication, and IV) the ability of an IVA to demonstrate physiotherapeutic exercises to

[4] https://developers.meta.com/horizon/documentation/unity/move-overview/

[5] https://developers.meta.com/horizon/downloads/package/oculus-lipsync-unity/

[6] https://github.com/Balint-H/modular-agents

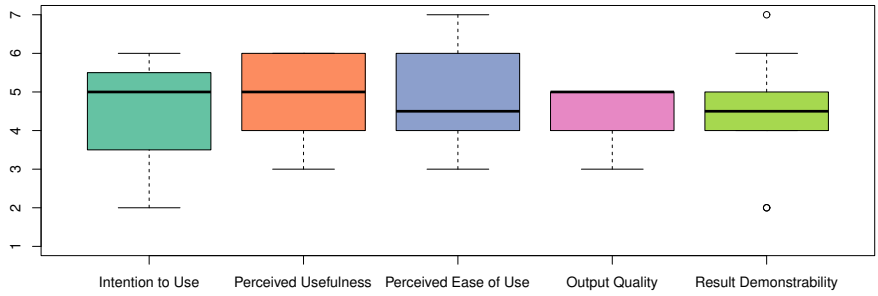

Figure 3: Mean values relating to Extended Technology Acceptance Model

users. We used an 11-point Likert scale to capture their prior experience with Unity development and virtual reality before using our tool. We also captured their experience with our IVH toolkit via standard questionnaires and open-ended questions. Four students (three men and one woman) with an average age of 27 ($SD = 6.93$) participated in our survey. One student entered 91 for their age and thus, was excluded from the mean calculation of the age. We incorporated the responses from all four participants in the evaluation of the remaining questions and questionnaires. They rated their prior experience with Unity development on average 6.75 (SD=2.5) and their prior experience with VR on average 4.5 (SD=1.29). The rest of this section reports on the results of this preliminary evaluation using standard usability and user experience questionnaires and open-ended questions.

System Usability Scale (SUS) [5]: The first questionnaire that we used was SUS which has 10 items and measures the usability of a system on a scale of 0-100. Our participants' responses to this questionnaire gave us an average SUS score of 56.88 ($SD = 6.88$). This indicates an OK to Good usability and a marginal acceptability score.

AttrakDiff [17]: The second questionnaire was AttrakDiff which consists of 28 items grouped into four sub-scales: Pragmatic Quality (PQ), Hedonic Quality relating to Identity (HQ-I), Hedonic Quality concerning Stimulation (HQ-S), and overall Attractiveness (ATT). This questionnaire helps evaluate overall user satisfaction and enjoyment. It has been also used in previous XR research for evaluating UX with XR systems [18]. The results show neutral to positive evaluations for all four sub-scales: Pragmatic Quality ($M = .89, SD = .69$), Hedonic Quality-Identity ($M = .71, SD = .57$), Hedonic Quality-Stimulation ($M = .39, SD = .47$), and Attractiveness ($M = .46, SD = .36$). Figure 2 shows the mean values for all four sub-scales.

Extended Technology Acceptance Model (TAM2) [46]: The next questionnaire was TAM2 which was employed to evaluate user acceptance and resistance to technologies. We used five sub-scales of this questionnaire to measure Intention to Use, Perceived Usefulness, Perceived Ease of Use, Output Quality (e.g., "The quality of the output I get from the system is high."), Result Demonstrability (e.g., "I have no difficulty telling others about the results of using the system."). As depicted in Figure 3, the evaluation of all scales was neutral to positive: Intention to Use ($M = 4.5, SD = 1.41$), Perceived Usefulness ($M = 4.81, SD = 1.28$), Perceived Ease of Use ($M = 4.88, SD = 1.15$), Output Quality ($M = 4.5, SD = .76$), Result Demonstrability ($M = 4.44, SD = 1.26$).

Open-ended questions: At the end of the survey, we asked several open-ended questions to capture further views of the participants about their use of our toolkit. The first question asked about their general experience with the toolkit. Two participants wrote that they used it for their Master's project, one of them also for their bachelor thesis and they were happy about it. Another user wrote, "The intelligent virtual agent is currently missing some controls, such as the left arm up and down, which is a little frustrating, but otherwise, it's easy to use".

The next question asked which part of the toolkit they liked or disliked. One user wrote, "I liked the fundamental functionality of the agent as these worked fairly well, the only problem is that the sound wasn't always recognized". Another user wrote that they liked that "many parts of the toolkit are controllable", while another user liked "the 3D Animation world and disliked the usability of the toolkit".

We also asked what they would like to see added and their answers included "more models", "more styles, hair, clothes for the avatar", and "higher usability when animating and building objects". We did not receive any responses regarding our question about what they wish to be removed or changed from the toolkit.

To our question of whether they would work with the toolkit again, we received three answers and all were a definite yes. Two users also wrote that they would recommend our toolkit to others, one of them mentioned the recommendation would be for specific use cases. Another user wrote a "yes and no" (maybe) response to this question.

5 Discussion and Outlook

In this work, we presented a preliminary implementation of our Intelligent Virtual Human Toolkit which consists of five modules each of which contributes to producing realistic human-to-human and human-to-agent interactions in XR.

The first module provides a set of diverse virtual humanoid 3D models in terms of represented gender, age, and ethnicity, to be used out of the box for multiple use cases including professional collaboration, manufacturing training, health care, and cultural heritage. The current set of models has several limitations which will be addressed by diversifying the body shape and size, increasing visual fidelity, improving rendering conditions - scene, lights, and shaders, and improving body deformation for more complex muscular movements like arm twisting or secondary shoulder/clavicle motion. Furthermore, this module will deliver a pipeline for creating humanoid 3D models based on users' photographs.

The second module facilitates detecting and classifying the actions of virtual humans in the virtual world in realtime. Using computer vision AI-based techniques, this module can already detect the action when a virtual human raises a hand in realtime. This action will be integrated in the future to provide a multimodal interaction between users and agents in an exemplary scenario described in Section Multimodal Interaction. In addition, this module will be further improved to include more actions and visual representations

of virtual humans such as the ones being holoported in realtime using 3D reconstruction techniques.

The third module provides speech and facial interactions for SAs and IVAs using AI-based services (such as speech-to-text, LLM, and text-to-speech). This allows for contextually relevant and emotionally intelligent dialogues between users and IVAs. Future work will make the IVAs visually intelligent to further facilitate vision-based communication between users and agents.

The techniques for interactive character animation that we are exploring are typically used in robotics and in physics simulation engines. They therefore use ragdoll-like characters, made of rigid or soft bodies assembled with joints that are actuated. However, these techniques are rarely used with skinned characters. In turn, both video games and virtual reality users use systematically skinned characters. It is an open question if these techniques can render the quality of movement that is expected when we compare these with more traditional interactive character animation techniques (typically, kinematic techniques) that are *de facto* considered industry standards. We plan to study in detail whether this is the case as a complement to our development efforts. An additional implicit assumption of our efforts in Body Animation and Interaction is that adopting physics-based interactive character animation techniques will help bring more life-like movement and dynamics to IVAs. This is an assumption that we plan to evaluate in terms of comfort and plausibility of the VR environment, as perceived by VR users [40].

Finally, the results of our preliminary evaluation with four participants indicated an OK to Good usability and a marginal acceptability score. The poor usability was also mentioned in the additional comments of the participants which needs to be improved in the future. Our toolkit at the time of evaluation contained one humanoid 3D model from Module I (which represented both an SA and an IVA) with one idle animation and basic speech-based interaction between the user and the IVA from Module III (i.e., facial and emotional expressions were not included). As a result, participants wished to see more models and ready-to-use animations included in the toolkit. This has been partially addressed in terms of ready-to-use 3D models for specific use cases included in Module I and will be further improved in the future by including a pipeline for creating 3D models from simple camera photos and videos by the developer users themselves. We also observed neutral to positive evaluations for all sub-scales of AttrakDiff and TAM2 questionnaires. This means that improvements need to be made to both the task-oriented and hedonistic qualities of our toolkit. Further comments from the participants revealed that despite all limitations, they would work with the toolkit again and would also recommend it to others.

We will continue improving our toolkit to make it more usable for creating XR solutions featuring IVHs. For future research, we will conduct several experiments to study the effects of interaction with single or multiple IVHs on users. For instance, we will study the effects of multi-modal interaction with IVHs in various XR scenarios including cultural heritage where IVHs represent tour guides and embody tourists, a manufacturing training scenario where both trainers and trainees are embodied as SA and receive assistance from an IVA, a health care scenario where IVHs help in reduction of medical procedure anxiety, and a professional collaboration scenario where users used IVHs to remotely participate in meetings in metaverse.

Acknowledgments

Early drafts of this submission were re-phrased for clarity using ChatGPT-4o. This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101135025, PRESENCE project.

References

[1] A. Aristidou, Q. Zeng, E. Stavrakis, K. Yin, D. Cohen-Or, Y. Chrysanthou, and B. Chen. Emotion control of unstructured dance movements. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pp. 1–10, 2017. doi: 10.1145/3099564.3099562

[2] L. Ascone, K. Ney, F. Mostajeran, F. Steinicke, S. Moritz, J. Gallinat, and S. Kühn. Virtual reality for individuals with occasional paranoid thoughts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020. 1

[3] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators & Virtual Environments*, 10(6):583–598, 2001. 2

[4] K. Bergamin, S. Clavet, D. Holden, and J. R. Forbes. Drecon: data-driven responsive control of physics-based characters. *ACM Transactions On Graphics (TOG)*, 38(6):1–11, 2019. doi: 10.1145/3355089.335653 2

[5] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996. 4

[6] J. S. Casaneuva. Presence and co-presence in collaborative virtual environments. 2001. 1

[7] H. Chu, S. Ma, F. De la Torre, S. Fidler, and Y. Sheikh. Expressive telepresence via modular codec avatars. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 330–345. Springer, 2020. 2

[8] H. G. Debarba, S. Chagué, and C. Charbonnier. On the plausibility of virtual body animation features in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(4):1880–1893, 2020. doi: 10.1109/TVCG.2020.3025175 2

[9] M. Dias, P. Coelho, R. Figueiredo, R. Carvalho, V. Orvalho, and A. Roche. Creating infinite characters from a single template: How automation may give super powers to 3d artists. In *ACM SIGGRAPH 2024 Talks*, pp. 1–2. 2024. 2

[10] M. Dias, A. Roche, M. Fernandes, and V. Orvalho. High-fidelity facial reconstruction from a single photo using photo-realistic rendering. In *ACM SIGGRAPH 2022 Talks*, pp. 1–2. 2022. 1, 2

[11] P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 3

[12] G. Freeman and D. Maloney. Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on human-computer interaction*, 4(CSCW3):1–27, 2021. 2

[13] J. P. Freiwald, S. Schmidt, B. E. Riecke, and F. Steinicke. The continuity of locomotion: Rethinking conventions for locomotion and its visualization in shared virtual reality spaces. *ACM Transactions on Graphics (TOG)*, 41(6):1–14, 2022. 2

[14] Y. Guo. A survey on methods and theories of quantized neural networks. *ArXiv*, 2018. doi: /10.48550/arXiv.1808.04752 3

[15] R. Hammady, M. Ma, C. Strathern, and M. Mohamad. Design and development of a spatial mixed reality touring guide to the egyptian museum. *Multimedia Tools and Applications*, 79(5):3465–3494, 2020. 1

[16] M. Hassan, Y. Guo, T. Wang, M. Black, S. Fidler, and X. B. Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–9, 2023. doi: 10.1145/3588432.359152 2

[17] M. Hassenzahl, M. Burmester, and F. Koller. Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. *Mensch & Computer 2003: Interaktion in Bewegung*, pp. 187–196, 2003. 4

[18] R. Horst and R. Dörner. Virtual reality forge: Pattern-oriented authoring of virtual reality nuggets. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–12, 2019. 4

[19] K. Huang and W. Gao. Real-time neural network inference on extremely weak devices: agile offloading with explainable ai. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, MobiCom '22, p. 200–213. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/

3495243.3560551 3

[20] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. 2002. 2

[21] L. Kruse, J. Hertel, F. Mostajeran, S. Schmidt, and F. Steinicke. Would you go to a virtual doctor? a systematic literature review on user preferences for embodied virtual agents in healthcare. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 672–682. IEEE, 2023. doi: 10.1109/ISMAR59233.2023.00082 1

[22] L. Kruse, F. Mostajeran, and F. Steinicke. The influence of virtual agent visibility in virtual reality cognitive training. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, pp. 1–9, 2023. 2

[23] C.-M. Kuo, S.-H. Lai, and M. Sarkis. A compact deep learning model for robust facial expression recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2202–22028, 2018. doi: 10.1109/CVPRW.2018.00286 2

[24] S. Lee, S. Lee, Y. Lee, and J. Lee. Learning a family of motor skills from a single motion clip. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. doi: 10.1145/3450626.34597 2

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, 2019. doi: 10.48550/arXiv.1907.11692 2

[26] J. Llobera and C. Charbonnier. Physics-based character animation and human motor control. *Physics of Life Reviews*, 2023. doi: 10.1016/j.plrev.2023.06.012 2

[27] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 2

[28] J. M. Markel, S. G. Opferman, J. A. Landay, and C. Piech. Gpteach: Interactive ta training with gpt-based students. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 2023. doi: 10.1145/3573051.3593393 2

[29] A. Maselli and M. Slater. The building blocks of the full body ownership illusion. *Frontiers in human neuroscience*, 7:83, 2013. 2

[30] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012. 2

[31] F. Mostajeran, M. B. Balci, F. Steinicke, S. Kühn, and J. Gallinat. The effects of virtual audience size on social anxiety during public speaking. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 303–312. IEEE, 2020. 2

[32] F. Mostajeran, N. Burke, N. Ertugrul, K. Hildebrandt, J. Matov, N. Tapie, W. G. Zittel, P. Reisewitz, and F. Steinicke. Anthropomorphism of virtual agents and human cognitive performance in augmented reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 329–332. IEEE, 2022. 2

[33] F. Mostajeran, N. Katzakis, O. Ariza, J. P. Freiwald, and F. Steinicke. Welcoming a holographic virtual coach for balance training at home: two focus groups with older adults. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1465–1470. IEEE, 2019. 2

[34] F. Mostajeran, P. Reisewitz, and F. Steinicke. Social facilitation and inhibition in augmented reality: performing motor and cognitive tasks in the presence of a virtual agent. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 323–328. IEEE, 2022. 2

[35] F. Mostajeran, F. Steinicke, O. J. Ariza Nunez, D. Gatsios, and D. Fotiadis. Augmented reality for older adults: exploring acceptability of virtual coaches for home-based balance training in an aging population. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020. 2

[36] S. Mystakidis. Metaverse. *Encyclopedia*, 2(1):486–497, 2022. 1

[37] S. Park, H. Ryu, S. Lee, S. Lee, and J. Lee. Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics (TOG)*, 38(6):1–11, 2019. doi: 10.1145/3355089.335650 2

[38] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. doi: 10.1145/3197517.320131 2

[39] S. Rings, S. Schmidt, J. Janßen, N. Lehmann-Willenbrock, and F. Steinicke. Empathy in Virtual Agents: How Emotional Expressions can Influence User Perception. In S. Hasegawa, N. Sakata, and V. Sundstedt, eds., *ICAT-EGVE 2024 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association, 2024. doi: 10.2312/egve.20241361 2

[40] M. Slater, D. Banakou, A. Beacco, J. Gallego, F. Macia-Varela, and R. Oliva. A separate reality: An update on place illusion and plausibility in virtual reality. *Frontiers in virtual reality*, 3:914392, 2022. doi: 10.3389/frvir.2022.914392 5

[41] H. J. Smith, C. Cao, M. Neff, and Y. Wang. Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2(2):1–17, 2019. doi: 10.1145/334025 2

[42] S. Starke, I. Mason, and T. Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. doi: 10.1145/3528223.3530178 2

[43] S. Starke, H. Zhang, T. Komura, and J. Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6):178, 2019. doi: 10.1145/3355089.3356505 2

[44] A. Steed, S. Frlston, M. M. Lopez, J. Drummond, Y. Pan, and D. Swapp. An 'in the wild'experiment on presence and embodiment using consumer virtual reality equipment. *IEEE transactions on visualization and computer graphics*, 22(4):1406–1414, 2016. 2

[45] H. Van Welbergen, B. J. Van Basten, A. Egges, Z. M. Ruttkay, and M. H. Overmars. Real time animation of virtual humans: a trade-off between naturalness and control. In *Computer Graphics Forum*, vol. 29, pp. 2530–2554. Wiley Online Library, 2010. doi: 10.1111/j.1467-8659.2010.01822.x 2

[46] V. Venkatesh and F. D. Davis. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2):186–204, 2000. 4

[47] 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. 1

[48] J. M. Wang, D. J. Fleet, and A. Hertzmann. Optimizing walking controllers for uncertain inputs and environments. *ACM Transactions on Graphics (TOG)*, 29(4):1–8, 2010. doi: 10.1145/1778765.1778810 2

[49] Y. Wu, Y. Wang, S. Jung, S. Hoermann, and R. W. Lindeman. Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and attraction. *Frontiers in Virtual Reality*, 2:641296, 2021. doi: /10.3389/frvir.2021.641296 1

[50] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, and W. Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 547–556. IEEE, 2019. 1

[51] H. Zhang, S. Starke, T. Komura, and J. Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018. doi: 10.1145/3197517.320136 2

[52] Y. Zhang, D. Gopinath, Y. Ye, J. Hodgins, G. Turk, and J. Won. Simulation and retargeting of complex multi-character interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023. doi: 10.1145/3588432.359149 2

[53] Y. Zhou, T. Li, B. Li, G. Wu, X. Meng, J. Guo, N. Wan, J. Zhu, S. Li, W. Song, C. Su, N. Chen, Y. Xing, Q. Wang, Y. Lin, and R. Li. Research on intelligent manufacturing training system based on industrial metaverse. In F. Hassan, N. Sunar, M. A. Mohd Basri, M. S. A. Mahmud, M. H. I. Ishak, and M. S. Mohamed Ali, eds., *Methods and Applications for Modeling and Simulation of Complex Systems*, pp. 28–43. Springer Nature Singapore, Singapore, 2024. doi: /10.1007/978-981-99-7240-1_3 1