AI-based Interactive Digital Assistants for Virtual Reality in Educational Contexts

Arelia McKern
Information Management in Engineering
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: uvwdd@student.kit.edu

Anjela Mayer
Information Management in Engineering
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: anjela.mayer@kit.edu

Lucas Greif
Information Management in Engineering
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: lucas.greif@kit.edu

Jean-Rémy Chardonnet
Arts et Metiers Institute of Technology
LISPEN, HESAM Université
Chalon-sur-Saône, France
Email: jean-remy.chardonnet@ensam.eu

Jivka Ovtcharova
Information Management in Engineering
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: jivka.ovtcharova@kit.edu

***Abstract*—This study presents and investigates a novel use of Artificial Intelligence (AI) for the creation of digital assistants in Virtual Reality (VR) environments for training and educational contexts. The concept proposed in this work couples two separate AI systems, the Movement Model (MM) and the Language Model (LM), to generate both conversational and visual responses simultaneously. The LM uses a two channel system to simplify communication with the MM, such that it responds to the user through one channel, and provides input to the MM through the other. This input allows the MM to produce context-relevant movements to assist communication with the user. Additionally, since this process is handled automatically by AI systems, it can provide a uniquely simple method for the creation and customization of digital assistants, with minimal technical knowledge or time investment from the user. A proof-of-concept prototype was implemented and underwent preliminary validation via a comparison to the Avatar Replay System (ARS) from our previous work in a user study. It was found that the AI assistants were able to interact with the users comparably to human recordings captured with the ARS assistants.**

I. Introduction

Every aspect of our lives is influenced by Artificial Intelligence (AI). It is estimated that 70 % of the global companies are preparing to integrate AI into their operations [1], [2]. In the rapidly changing realm of AI, Large Language Models (LLMs) offer a spectrum of possible changes, reshaping procedures in a variety of industries.

This study investigates an innovative use of AI for the development of digital assistants in virtual reality environments. The creation of new Non-Player Characters (NPCs) has traditionally been a time-consuming and costly process, necessitating specialized knowledge. This study presents an approach to creating NPCs solely with AI technologies, employing an AI language model for communication and providing input to an AI motion model to produce coordinated movements. This dual-purpose application aims to establish a more vibrant and reactive virtual entity.

In education, these interactive NPCs can profoundly transform how instructional content is delivered, serving as customizable virtual tutors and assistants that dynamically adapt to student interactions. Such adaptability has the potential to significantly increase engagement and facilitate personalized learning experiences.

This paper presents a coupled AI system that seamlessly integrates a language model (LM) for communication with a motion model (MM) to generate NPC movements. This innovative synergy not only facilitates the creation of NPCs that interact more naturally and in real time with users, but also simplifies the development process, making it accessible to non-programmers. Such NPCs can dynamically adapt to interactions, providing customized tutoring and assistance in educational settings. This adaptability potentially increases engagement and personalizes the learning experience, marking a significant advancement in educational technology.

Following this introduction, Section II presents a review of related work that sets the context and outlines previous methodologies. Our concept, explaining the integration of the two AI models is outlined in Section III. Following this, the prototype implementation is summarized in Section IV and the preliminary results are presented in Section V, discussing their implications. Finally, Section VI concludes with a summary of our findings and suggestions for future research directions.

II. Related Work

In this section, a review of work related to this paper is presented, with each subsection focusing on a single thematic area. Subsection II-A details the movement generation solution used in this paper. Following this, Subsection II-B elaborates upon works investigating the perception of NPCs, and uses those results to make predictions about the usefulness of the approach presented in this paper. Finally, Subsection II-C explains the related previous work by the authors of this paper.

A. Body movement generation

A 2022 study aimed to address the problem of automated human motion generation from text with a two-phase approach, where the length of the movement is determined separately from the movement itself [3]. A new dataset of captioned human motions was created for this study, "HumanML3D", which was tested alongside the main dataset relied upon in prior work, "KIT Motion-Language". This test was said to "demonstrate the superior performance of our approach over existing methods" with empirical evaluations [3].

Another study published later in the same year proposed the use of a diffusion model for movement generation [4]. The study compared the generated motion outputs to contemporary methods, including to [3], which it is reported to have outperformed slightly in both quantitative and qualitative tests. For the purposes of an initial proof of concept however, the methods' reported output quality was similar enough that it's use as a selection criterion for the model was considered arbitrary in this case. In this case, the selection of [3] over [4] was made qualitatively based on ease of implementation into the test system.

Further existing motion generation algorithms such as Language2Pose [5] or Text2Gesture [6] were not considered for use in this case due to the reported quality differences in both of the 2022 papers [3] and [4].

B. Virtual Assistants

The impact of the embodiment of the assistant on the user experience, both for human and artificial assistants, in the field of patient care was investigated in [7]. The embodiment of the assistant was shown to have a significant impact on the engagement of the patient and the willingness to use the assistant, as well as the social richness and social presence of the assistant. The overall results demonstrated that human assistants were overall higher rated than digital assistants, but also that embodied assistants were rated more highly than disembodied assistants.

Furthermore, a 2020 study expanded on this by investigating the effects of virtual assistant[1] embodied in collaborative decision making. Here, it was found that the presence of a virtual assistant in any form has a positive effect on task performance compared to the control case with no assistance, and that the embodiment of the assistant reduced the perceived task load compared to the disembodied assistant.

The positive effects of the digital assistant embodiment provide a theoretical basis for the idea that an AI assistant controlling its own embodiment could be beneficial.

In 2022 a study was published that investigated the effect of interacting with a digital assistant through natural language, instead of a standard graphical user interface [9]. In this study, communication through natural language was found to improve naturalness, authenticity, realism, and fluency of the interaction with the digital assistant, making it more likeable, and making it appear more energetic and active.

[1]Cambridge dictionary defines virtual assistants the same way as digital assistants: https://dictionary.cambridge.org/dictionary/english/virtual-assistant. As such, even though [8] specifically uses the term virtual assistant, the results are considered relevant to digital assistants as well.

C. Previous Work

The Avatar Replay System (ARS) that uses mixed reality (MR) technologies has previously been developed to capture and replay human operations in space [10], [11]. This system uses MR device tracking, including HoloLens 2 and Oculus Quest (1 and 2) Head Mounted Displays (HMDs), to record head and hand tracking information of the HMD user. During the replay, these captured movements were used to generate full-body animations of a virtual character, using the inverse kinematic capabilities of the Oculus XR plugin.

This application of MR technologies to manually record and replay human movements serves as a critical precursor to the automated processes developed and investigated in this work that utilize AI.

III. Concept

The basis of the proposed method for developing digital assistants is based on two fundamental types of AI system: the Movement Model (MM) and Language Model (LM). Although each model is responsible for the respective portion of generating an NPC, simply allowing the two to act completely independently of each other would likely create a fractured behaviour output of the NPC. As such, aiming to improve the cohesiveness of the NPC behaviour, a novel interaction between the two is proposed, which has not been investigated in prior works.

This approach begins with a separation of functions within the LM, where two separate channels are opened with different initial system messages describing their function, but the same message describing the scenario in which the user is in. The user's input is fed into both of these channels at the same time, but the user only receives a response from one channel, while the other channel passes its output on to the MM, as can be seen illustrated in the flowchart in Figure 1.

The MM then accepts the outputted message from the LM and uses it to generate motions; therefore, the implemented MM must be able to generate motions from text. Once the movements have been generated, they are sent to the VR environment to be displayed to the user, ideally played simultaneously with the Text-to-Speech (TTS) audio generated from the response to the user by the LM. The described interactions can be seen visualized in an interaction diagram in Figure 2.

Creating a new NPC with this system should be as simple as changing the initial scenario-describing message that both channels of the LM receive when starting, and since both the motions and language responses are generated live for the user, the resulting NPC should also be interactive and able to react to any potential user input. Both of these qualities contrast starkly with prior methods of NPC creation, which can often be a time intensive process which requires a level of expertise in the creation method to do well, and produces NPCs with limited intractability, only able to react dynamically to user

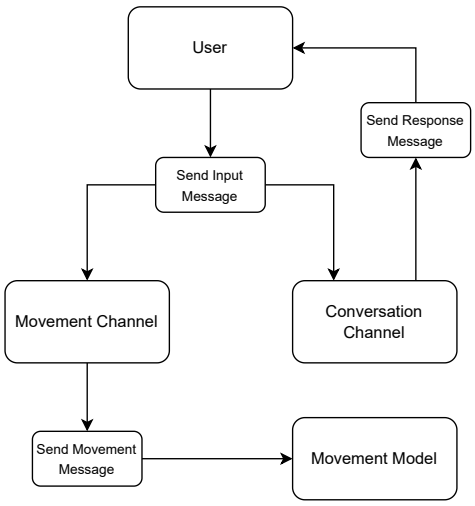


Fig. 1: Channel flowchart

inputs that the creator of the NPC was able to foresee and prepare for.

IV. Prototype Implementation

In order to transition from the concept to a full prototype, it was primarily necessary to choose the specific programs and existing code bases to draw from, and implement the necessary interfaces. The process started with the choice of "gpt-3.5-turbo-0125"[2] as the LM. The decision to use this particular model was based on its cost-effectiveness and reliability, which provided a viable basis for initial testing. Importantly, if this model proves effective in our application, it suggests that more advanced models could also be successfully integrated in the future. The model implemented in [3] was chosen for the MM, and continued with the implementation of the interface, and slight modifications to the code base from [3].

[2]version dated: 31.03.2024

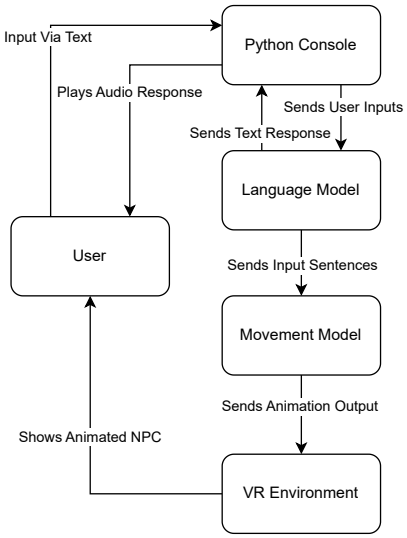


Fig. 2: Interaction diagram

The python[3] code for the project consists primarily of the MM code, with the interface to ChatGPT integrated via OpenAIs python API directly into the existing main file. Instead of generating movements from existing text files as originally implemented in [3], a main loop is created where the movement generation is repeated each time the user interacts with the LM, and the text input for the MM generation is taken from the output of the movement conversation by the LM.

In the case of the prototype, the output from the MM is then transferred to the VR environment, implemented here in Unreal Engine[4] 5.3.2, where the movements are displayed to the user. The prototype displays the movements to the user using visible spheres representing each of the points of the generated output model, animated with each frame of the output movement corresponding to one frame in the Unreal Engine output.

At the same time, the response from the LM is played, such that the user is able to receive both visual and auditory feedback. This response is played simultaneously with the animations, but not synchronously, since the scope of this work was limited and ensuring synchronous playback was deemed outside the scope.

The implemented prototype as described can be found on GitHub[5].

V. Preliminary Results and Discussion

The implemented prototype was validated and compared with the ARS system from our previous work. The purpose of this pilot study was to investigate the following research question: *Is the AI powered character animation perceived similarly to animations created with the ARS?*

A. Preliminary study results

For validation, a within-subject user study was conducted with $N = 9$ participants ($N = 4$ female, $N = 5$ male), testing the following conditions:

- $Movement_{AI}$: The movement was generated by prompting the AI prototype
- $Movement_{ARS}$: The movement was generated by recording it with the ARS system

The subjects therefore were wearing the Oculus Quest 1 Virtual Reality (VR) HMD visualizing a set of animations in front of them within a virtual environment. Initially, the subjects were asked to rate the ease with which they were able to understand what each NPC was communicating, as well as how useful the NPCs would be as learning tools, using a 7-point Likert scale with 1 being *"very easy"* or *"very useful"* to 7 *"very difficult"* or *"not useful at all"*. This was measured for two different learning scenarios, with users testing both

[3]Python version 3.7.9, chosen because it was the existing basis for the MM code base.

[4]Unreal Engine: https://www.unrealengine.com/de

[5]AI NPC Creation: https://github.com/AreliaEmber/AI_NPC_Creation/

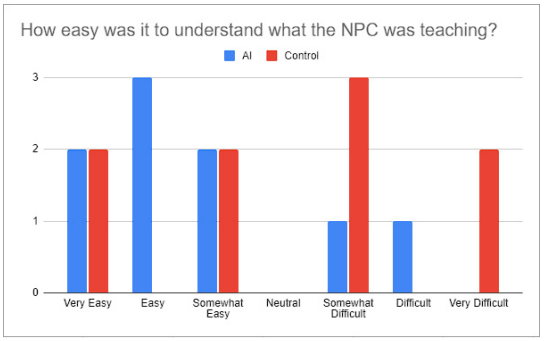


Fig. 3: Chart showing a comparison of how understandable the users found each NPC.

TABLE I: Validated Movements of $Scenario_{Dance}$ by Condition

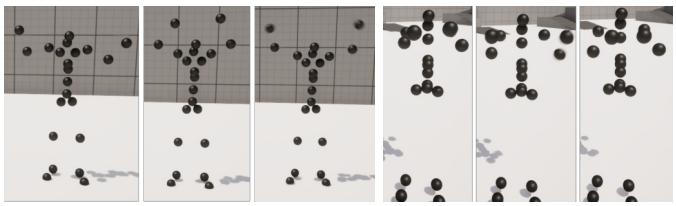
Condition	Movement
$Movement_{AI}$	Step to the left Jump Wave both arms
$Movement_{ARS}$	Point forward Turn 360 degrees on the spot Wave both arms

scenarios with both conditions. The scenarios tested were as follows:

- $Scenario_{Pythagoras}$: The NPC taught the basics of Pythagorean theorem
- $Scenario_{Dance}$: The NPC taught the user a very simple dance

It was observed that the AI system was consistently rated equal to or better than the control system when considering the whole dataset. In particular, the AI NPC was rated higher on understandability ($M = 2.8$, $SD = 1.7$) than the ARS NPC ($M = 4.1$, $SD = 2.3$), and approximately equally on usefulness as a learning tool (AI: ($M = 3.2$, $SD = 1.6$), ARS: ($M = 3.3$, $SD = 1.7$)). Figure 3 illustrates graphically the user rating of the understandability of each NPC.

After completion of the interaction tasks with each condition, participants were asked to identify movements from a list of descriptions. Table I lists the validated movements by condition and Figure 4 depicts the view of the *wave both arms* movement for each condition.



(a) AI condition (b) ARS condition

Fig. 4: Sequence of the wave movement.

B. Discussion

Users were more frequently able to accurately identify the AI-generated movements, and consistently rated the movements as easier to identify than the ARS movements. It could be argued that this is a result of user error on the part of the user recording the initial movements in ARS, for example that the user that recorded the motion seen in Figure 4b recorded the motion of waving their hands instead of waving their arms, however, it could equally be argued that user error when setting up or interacting with an AI NPC is also a possibility.

The preliminary results indicate that the $Movement_{AI}$ condition can achieve at least similar perceived movement animations compared to $Movement_{ARS}$. Thus, AI-generated NPCs are comparable to motion tracking and maybe even outperform them in terms of creation efficiency, effectiveness, and interactivity.

Additionally, the higher ratings in the learning scenarios and the fact that more than half of the participants chose the AI NPC as the preferred NPC for learning and understandability indicate that the approach is viable and warrants further research to expand upon the prototype.

VI. Conclusion

AI is gaining a significant role in our daily and professional life. Virtual agents play an important role in education, which can significantly benefit from AI models in terms of interactivity and engagement. In this work, a concept was presented coupling two AI models to facilitate NPC animations, which can furthermore generate response-animations corresponding to the human interactions. A proof-of-concept prototype was implemented and tested with several users in a pilot experiment. The preliminary results indicate that the AI generated movement is at least as well perceived as motions captured from humans. In future work, this approach can be applied to NPC characters to facilitate intelligent agents in immersive virtual environments, such as virtual tutors or assistants, which can interact with humans in real-time. Furthermore, future research could focus on further validating our results and improving the system architecture. Such efforts could yield more reliable data to refine the model and broaden its potential uses, improving the seamlessness and engagement of virtual interactions.

Acknowledgment

The authors would like to thank the Ministry of Science, Research and Arts of the Federal State of Baden-Württemberg for the financial support of the projects within the InnovationsCampus Future Mobility (ICM).

References

[1] J. Bughin, J. Seong, J. Manyika, M. Chui, and R. Joshi, "Notes from the ai frontier: Modeling the impact of ai on the world economy," *McKinsey Global Institute*, vol. 4, 2018.

[2] J. Brasse, H. R. Broder, M. Förster, M. Klier, and I. Sigler, "Explainable artificial intelligence in information systems: A review of the status quo and future research directions," *Electronic Markets*, vol. 33, no. 1, p. 26, 2023.

[3] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," *Computer Vision and Pattern Recognition*, 2022.

[4] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *arXiv.org*, 2022.

[5] C. Ahuja and L.-P. Morency, "Language2pose: Natural language grounded pose forecasting," in *2019 International Conference on 3D Vision (3DV)*, 2019, pp. 719–728.

[6] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha, "Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021, pp. 1–10.

[7] K. Kim, N. Norouzi, T. Losekamp, G. Bruder, M. Anderson, G. Welch, and G. F. Welch, "Effects of patient care assistant embodiment and computer mediation on user experience," *International Conference on Artificial Intelligence and Virtual Reality*, 2019.

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

[9] K. Buchta, P. Wójcik, K. Nakonieczny, J. Janicka, D. Galuszka, R. Sterna, and M. Igras-Cybulska, "Modeling and optimizing the voice assistant behavior in virtual reality," *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2022.

[10] A. Mayer, T. Combe, J.-R. Chardonnet, and J. Ovtcharova, "Asynchronous manual work in mixed reality remote collaboration," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13446 LNCS, p. 17 – 33, 2022.

[11] A. Mayer, A. Rungeard, J.-R. Chardonnet, P. Häfner, and J. Ovtcharova, "Immersive hand instructions in ar – insights for asynchronous remote collaboration on spatio-temporal manual tasks," in *2023 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2023, pp. 1–6.