A Conversational Virtual Agent with Physics-based Interactive Behaviour

Joan Llobera* Artanim Foundation
Ke Li University of Hamburg
Pierre Nagorny Artanim Foundation
Caecilia Charbonnier Artanim Foundation
Frank Steinicke University of Hamburg

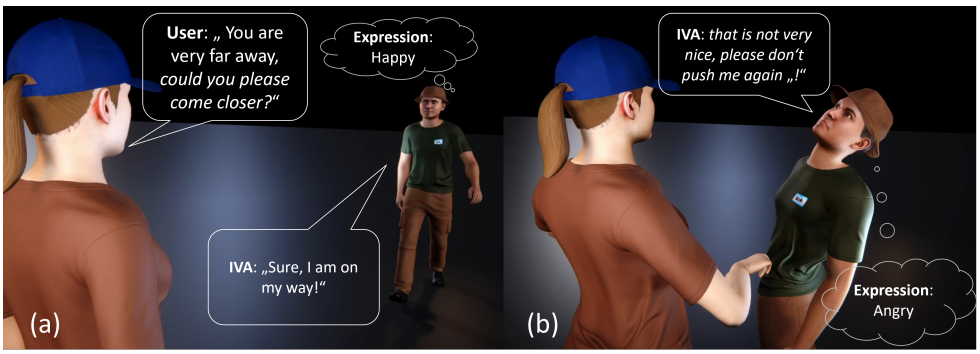

Figure 1: Illustration of physics-based interactive behaviour. (a) The user calls out to the IVA from a distance: "*You are very far away, could you come closer?*" The IVA responds with "*Sure, I am on my way!*" and approaches using a physics-based animation controller, maintaining a customizable social distance. (b) The user unexpectedly pushes the IVA, for example, when it enters the user's personal space. The IVA reacts by losing its balance and nearly falling to the ground. In response, it displays an angry facial expression, and the LLM generates a speech response: "*That is not very nice, please don't push me again!*"

Abstract

We demonstrate an Intelligent Virtual Agent (IVA) in VR with motion driven by a physics-based controller. This enables dynamic, unscripted behaviour like autonomously maintaining social distance. Moreover, our system links physical interactions, such as a user pushing the agent, to a Large Language Model (LLM). The LLM generates context-aware verbal responses to the physical event, bridging the gap between low-level motor control and high-level decision making. In the demo session, attendees can directly test the robustness of these integrated social and physical behaviour.

Index Terms: Intelligent Virtual Agent, Physics-based Animation

1 Introduction

A significant challenge in Virtual Reality (VR) is the creation of Interactive Virtual Agents (IVAs) that move and interact with physical believability. The core of this issue lies in a fundamental disconnect between IVA's cognition and motor control. Typically, an agent's high-level decision-making (*what to do*) is separate from its physical execution (*how to move*). This limitation leads to a reliance on a finite library of pre-authored animations [5]. As a result, most previous embodied IVAs generate behaviors using traditional animation databases, which often makes them appear rigid and robotic. Therefore, even if basic principles of interpersonal coordination have been validated with VR characters [6], we still don't have virtual characters that can spontaneously engage in interpersonal coordination, mimicry, or convincingly portray other nonverbal language in an interactive conversation.

Recent advances in artificial intelligence (AI) have demonstrated the remarkable ability of Large Language Models (LLMs) to simulate human-like conversations. However, these systems face a fundamental limitation: they are disembodied. Their intelligence is mainly based on statistical patterns in large text datasets, which separates them from the rich, multi-sensory inputs such as vision and touch that ground human language in physical experience. As a result, LLMs often struggle with reasoning skills on spatial or body-centered concepts [8] and typical conversational IVAs powered by LLMs alone cannot intuitively react to spatial or non-verbal cues such as users' gestures like looking away or pointing in a direction. However, these abilities are essential for effective human-AI collaboration.

In this demo, we present an interactive work-in-progress system that aims to bridge the gap between an IVA's LLM-based decision-making and its physical execution, enabling more engaging and co-present human-AI interactions. Our work builds on the multimodal agent developed by Li et al. [5], which featured high-level capabilities to see, hear, speak, and trigger pre-scripted actions based on LLM outputs. Here, we introduce a key addition: a physics-based controller that governs the agent's body motion in real time. For example, rather than relying on a fixed "*walk*" animation, our IVAs use a physics-driven animation controller trained with deep reinforcement learning and motion matching to simulate realistic walking behaviors of humans.

We demonstrate that novel interactive possibilities emerge naturally from the dynamic interplay between the IVAs and the user's actions in real time. As shown in Figure 1, these capabilities include not only maintaining a coherent, LLM-driven conversation but also engaging in physically plausible interactions, such as autonomously adjusting position to preserve social distances or realistically stumbling when pushed by the user while simultaneously displaying an "*angry*" facial expression and corresponding speech response. Through this demo, we aim to inspire future work that integrates low-level motor control with high-level LLM-based decision making in embodied IVAs, advancing research and technical development in convincing nonverbal communication [2] and supporting the creation of next-generation agents with truly embodied social and spatial reasoning skills for human-AI interactions [4]. This is, to our best knowledge, the first demo to couple emergent, physics-based reactions with LLM-driven intelligence.

*e-mail: joan.llobera@artanim.ch

2 System Implementation

This technical demo runs as a standalone Android build on the commercial VR headset Oculus Quest 3. It was developed using Unity 3D (version 2023.2.20).

2.1 Speech Interaction

We implement speech interaction between users and the IVA using functionalities from the intelligent virtual human toolkit developed by Li et al. [5]. The IVA can listen to the user's speech input via Google Cloud Speech-to-Text (STT). The transcribed text, combined with a system prompt that defines the IVA's role, is then processed by a LLM such as the OpenAI's GPT-4o or Google Gemini 2.5 pro model to generate responses. These responses include not only natural language replies but also high-level behavioral decisions, such as body movements, gaze shifts, or facial expressions, using structured output and function-calling capabilities. The generated text response is converted to audio using a text-to-speech (TTS) cloud service such as ElevenLabs or Google Cloud. In this demo, the IVA supports three high-level, physics-based behaviors: "*walk towards player*", "*step back*", and "*walk forward*". These actions allow the agent to dynamically adjust its proximity to the user based on the interaction context.

2.2 Physics-based Body controller

The body controller is responsible for navigating the virtual space, enabling the IVA to approach the user in VR while maintaining an appropriate social distance. This includes walking forward, side steps, back steps, and idling behavior. Idling, in particular, is an aspect that is generally not adressed in the scientific literature on physics-based humanoid controllers.

Our controller is based on a reimplementation of DReCon [1], a state-based physics controller in which a motor policy is trained using deep reinforcement learning to maintain physical balance while matching target poses generated at each frame by Motion Matching, a kinematic algorithm for navigation. We use MxM for motion matching, an off-the-shelf Unity library[1]. The motion data used was captured and processed by the Artanim Foundation. For physical simulation, we use MuJoCo [7] [2]. To make it compatible with Android-based VR headsets, we created a custom build integrated into a Unity3D package [3]. The physics-based controller was implemented using the Modular Agents framework[4], a custom open-source library and set of example scenes built on Unity's ML-Agents reinforcement learning toolkit.

The training environment was designed to combine different walking and human-IVA interactions. These included moments where the IVA approaches a VR user proxy, maintains a socially appropriate distance while the proxy moves in random directions, and walks to various spatial targets. Regular physical perturbations were introduced to improve the controller's robustness. In this demo, the trained controller is placed into the VR scene, where the VR user proxy is replaced by the actual position of the VR user. The target position of the IVA relative to the user is modified dynamically using functions triggered by keywords.

2.3 Facial Expression Controller

Traditional keyframe animation for facial expressions is incompatible with our physics-based character controller. To address this, we implemented a facial animation system using blendshape presets based on the Facial Action Coding System (FACS) [3]. This system allows for the real-time display of six primary emotions (sadness, happiness, fear, anger, surprise, and disgust), each with a controllable intensity level from 0 to 1. All of the facial expression animations are pre-defined blendshape presets can be triggered by LLM/VLM structured outputs. For real-time lip sync, we integrate the Oculus OVR Lip Sync API, which enables real-time lip synchronization by mapping spoken audio to 15 predefined viseme blendshapes.

[1] https://assetstore.unity.com/packages/tools/animation/motion-matching-for-unity-145624

[2] now open source at: https://mujoco.org/

[3] https://github.com/joanllobera/mujoco-bin

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

3 Conclusion and Outlook

In this work, we demonstrated the feasibility of running a conversational IVA with physics-based interactive behaviors on a standalone, commercially available VR headset. This system represents an initial step toward building more embodied, responsive, and socially aware virtual agents. In future work, we plan to extend the system to support richer interactions with physical elements such as cutlery, tools, furniture, and obstacles. We also aim to enhance nonverbal communication through expressive gestures, gaze, and body language, and to further develop the IVA's ability to process multisensory information in conjunction with physics-based animation. Our long-term goal is to enable VR characters to understand and respond to complex physical and social scenarios, such as saying, "*I can't pick up the book because you locked it in the safe.*"

Acknowledgments

The 3D virtual human are provided by Didimo [5]. Valerie Juillard and Yves Schmid Dornbierer (both from Artanim) took care, respectively, of retargeting the motion captured data to the targeted character, and adjusting the motion matching system to use it. Yves Schmid Dornbierer was also instrumental in compiling MuJoCo for Android. This work was supported by the European project Presence XR, a Horizon Europe Innovation Project co-financed by the EC under Grant Agreement ID: 101135025 and SEFRI under Grant Contract 23.00466. Some parts of the early drafts of this submission were rephrased for clarity and grammatical correctness using OpenAI's GPT-4o model and Google's Gemini.

References

[1] 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: 10.1145/3355089.3356536 2

[2] A. Deichler, S. Wang, S. Alexanderson, and J. Beskow. Learning to generate pointing gestures in situated embodied conversational agents. *Frontiers in Robotics and AI*, 10:1110534, 2023. 1

[3] P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2

[4] R. Krishna, D. Lee, L. Fei-Fei, and M. S. Bernstein. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39):e2115730119, 2022. doi: 10.1073/pnas.2115730119 1

[5] K. Li, F. Mostajeran, S. Rings, L. Kruse, S. Schmidt, M. Arz, E. Wolf, and F. Steinicke. I hear, see, speak & do: Bringing multimodal information processing to intelligent virtual agents for natural human-ai communication. *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 1648–1649, 2025. 1, 2

[6] J. Llobera, V. Jacquat, C. Calabrese, and C. Charbonnier. Playing the mirror game in virtual reality with an autonomous character. *Scientific Reports*, 12(1):21329, 2022. doi: 10.1038/s41598-022-25197-z 1

[7] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109 2

[8] Q. Xu, Y. Peng, S. A. Nastase, M. Chodorow, M. Wu, and P. Li. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, pp. 1–16, 2025. doi: 10.1038/s41562-025-02203-8 1

[5] https://www.didimo.co/