I Hear, See, Speak & Do: Bringing Multimodal Information Processing to Intelligent Virtual Agents for Natural Human-AI Communication

Ke Li[1a*], Fariba Mostajeran[1b*], Sebastian Rings[1], Lucie Kruse[1], Susanne Schmidt[2], Michael Arz[1], Erik Wolf[1], Frank Steinicke[1]

[1]Human-Computer Interaction, University of Hamburg, [2]HIT Lab NZ, University of Canterbury

Figure 1: Illustration of a VR user interacting with an embodied IVA using our system. The IVA autonomously responds to the user with realistic human-like behaviors, including speech and non-verbal cues such as body language and subtle facial expressions.

Abstract

In this demo paper, we present an Extended Reality (XR) framework providing a streamlined workflow for creating and interacting with intelligent virtual agents (IVAs) with multimodal information processing capabilities using commercially available artificial intelligence (AI) tools and cloud services such as large language and vision models. The system supports (i) the integration of high-quality, customizable virtual 3D human models for visual representations of IVAs and (ii) multimodal communication with generative AI-driven IVAs in immersive XR, featuring realistic human behavior simulations. Our demo showcases the enormous potential and vast design space of embodied IVAs for various XR applications.

Index Terms: Intelligent Virtual Agents, Extended Reality, Embodied AI, Human-AI Interaction

1 Introduction & Background

The latest breakthroughs in generative AI such as generative pre-trained transformer (GPT) and large language models (LLMs) have revolutionized the capabilities of IVAs in understanding complex text inputs and generating context-aware human-like text and audio responses [2]. Despite these breakthroughs, user interactions with IVAs remain largely confined to conventional chatbot-based interfaces, such as ChatGPT, limiting their potential to create more engaging and natural human-AI interactions. This raises a critical research question on how these foundational generative AI models can be best integrated into IVAs to enable richer human-AI interactions. For various immersive experiences, virtual humans are used to not only enable users to engage with AI-driven IVAs in social settings but also provide them with a self-representation [1]. These avatars aim to create a body-ownership illusion and a compelling sense of social and co-presence for users in immersive XR environments.

In this paper, we demonstrate an XR framework that enables streamlined workflows for creating and interacting with virtual humans, generating personalized avatars and customizable generative AI-driven IVAs. In the first use-case demo, we present workflows for creating high-quality personalized IVAs with realistic visual appearance for XR applications by replicating a user's facial features and voice using commercially available software and cloud services. In the second use-case demo, we present how users can engage in conversations with custom-made IVAs in XR using multimodal inputs such as speech, images, and system prompts. As shown in Figure 1, the IVA provides text-to-speech (TTS) responses and autonomously exhibits realistic, context-appropriate human behavior, including body language, facial expressions, and lip-sync for speech synthesis. Through these demos, we aim to explore the vast design space of IVAs, while building XR applications that facilitate engaging and natural human-AI interactions.

2 System & Workflows

Customizable IVA Creation Workflows We have established an IVA creation workflow to accommodate diverse user preferences and application requirements. This workflow enables the rapid integration of realistic humanoid 3D models into the Unity game engine. A 3D character can be generated from a 2D image using the Character Creator (CC4) Headshot plugin [1] and exported with ARKit-compatible facial blend shapes for expression animation. For user-controlled avatars in XR, the ARKit blend shape standard is also compatible with Meta's Movement SDK for facial expression tracking using the Meta Quest Pro headset. For voice replica-

*These authors contributed equally to the work.
Video demo: https://youtu.be/BTKCyC0GgXg
[a] e-mail: ke.li@uni-hamburg.de
[b]e-mail: fariba.mostajeran.gourtani@uni-hamburg.de

[1]https://www.reallusion.com/character-creator/headshot/

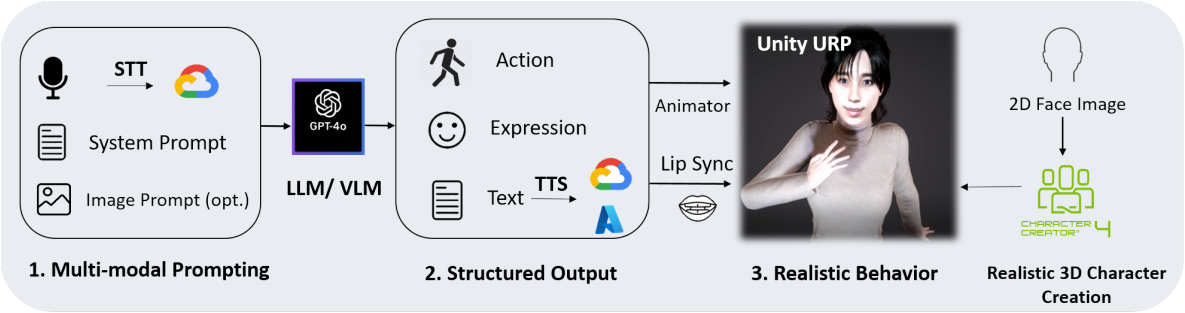

Figure 2: System overview in creating realistic human-like behaviors for the IVA, including 1. multi-modal prompting, combining text-based user message with system prompt that contains a list of possible agent actions and facial expressions, with optional image prompt as support, and 2, the structured output from the LLM/VLM, containing selected action, facial expression, and text response, and 3. realistic IVA behavior combining the multi-modal output, rendered in Unity Universal render pipeline for realistic agent appearance visualization.

tion, the system integrates customizable neural voice services from Microsoft Azure and ElevenLabs to enable real-time, cloud-based TTS with personalized voice. We developed a configuration script in Unity to streamline the setup by connecting the IVA to predefined animation controllers, AI services, and lip-sync tools in a single step. This workflow accelerates IVA creation, making them quickly deployable for various XR applications.

Multimodal Prompting As shown in Figure 2, our system is powered by state-of-the-art foundation models, including large language models (LLMs) and visual language models (VLMs) such as OpenAI's GPT-4o via its API, which can be exchanged with other advanced models, enhancing the tool's versatility. We use a multimodal prompting strategy, sending the user's message to the foundation models as a combination of text, system prompts, and 2D images. The text input, representing the user's message, is obtained via speech-to-text (STT) services like the Google Cloud API. An optional image prompt includes a 2D image captured from the agent's perspective, providing a visual context of the XR environment. The system prompt specifies the goal of the conversation and includes a list of potential actions and facial expressions that the LLM/VLM can trigger in its response. Facial expression animations are exported from the Digital Soul 100+ library (CC4) into Unity, while action animations are sourced from Adobe's Mixamo gesture pack and CC4's motion plus package.

Structure Output In our system, the LLM/VLM are prompted to generate structured outputs, responding to the user's query with both text and appropriate non-verbal behaviors, such as body language and facial expressions. The text response is converted to speech using TTS services from Google Cloud API, Azure Custom Voice API, or Eleven Labs TTS API. This audio is then synchronized with the agent's speech using the OVR Lip Sync API, which aligns the speech with a standard 1:1 viseme set and modifies 15 viseme blendshapes. Non-verbal cue animations (such as nodding) are triggered during the TTS processing, which also compensates for network latency during TTS generation. A single animator manages both the IVA's actions and facial expressions, while two distinct animation layers and animation masks separate the head animation from the body animation, ensuring coherent and synchronized movements.

Realistic Behavior Finally, realistic human-like behavior is simulated using the Unity Universal Render Pipeline (URP), which strikes a balance between rendering realism and performance optimization. As shown in Figure 1, our system enables the IVA to respond to user messages in various social contexts autonomously, enhancing the interaction beyond simple text exchanges with generative AI. For example, when receiving a compliment, the IVA will trigger a gentle smile, perform a happy dance, and provide speech feedback to express happiness.

3 Demonstration

In the demo session, we showcase our system and workflow using an Alienware laptop with an RTX 2080 GPU and a Meta Quest Pro headset. Participants can create a custom IVA by providing a 2D image of their face to use the CC4 headshot plugin. Afterward, our Unity framework can quickly configure the custom-made 3D character from CC4 as IVA. Our system enables them to perceive and interact with their personalized IVA in an immersive XR environment. Participants can customize the system prompt and choose the interaction modality for their intended conversations, allowing for various engaging human-AI interactions.

4 Conclusion

In this paper, we showcase an XR system and workflow that quickly brings a customized IVA into immersive XR applications, making the solution versatile and impactful for various use case scenarios and setting the foundation for future exploration of more complex human-AI interaction in XR. In future work, we aim to expand the design space of embodied IVAs in the Metaverse by enhancing system capabilities to support multi-agent interactions, more sophisticated perception-action loops, and additional interaction modalities, such as gaze behavior consistency and haptic feedback. In addition, we acknowledge the potential for users to misinterpret AI emotional responses and suggest further research into providing users with clearer cues about the AI's intent and the reasoning behind its emotional expressions.

Acknowledgments

Early draft of this paper was helped by OpenAI's GPT-4o model for improving writing clarity. This work has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101135025, PRESENCE project, and Germany's Ministry of Education and Research (BMBF), grant No 16SV8878, HIVAM project.

References

[1] K. Kim, L. Boelling, S. Haesler, J. N. Bailenson, G. Bruder, and G. Welch. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 105–114, 2018. 1

[2] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. 1