Virtual Character Interaction under the Fusion of Generative AI and XR: Evolution, Methods and Future

Guanhao Zhao

The High School Affiliated to Beijing Normal University, Beijing, China
tangmutomyyds@hotmail.com

Abstract. With the rapid development of generative AI and XR technology, virtual character interaction is entering a new era. Traditional characters are subject to preset scripts, which are difficult to meet the needs of users for deep emotional connection and personalized experience. Therefore, it is essential to better address these limitations and offer more engaging and customized experiences in the digital realm. This paper systematically summarizes the research background, core technology framework and technological breakthroughs in this field, and mainly analyzes the representative cases of emotional speech synthesis, large language models (LLMS) and physical real action generation. By integrating the research results of the preface, this paper proposes a framework integrating generative AI and XR technology, aiming to build a more realistic virtual role with deep emotional connection, thereby satisfying users' needs. At the end of the paper, the prospect of multimodal integration and personalized customization is discussed, and the challenges in technology, computing power and ethics are discussed.

Keywords: Generative artificial intelligence, Expand reality, Virtual role interaction, Multimodal interaction

1. Introduction

In various application scenarios such as digital entertainment, virtual social networking, online education and mental health, virtual roles have become an important way of human-computer interaction. However, users' expectations of virtual roles have changed from "functional NPC" to "emotional partner".

Take the well-known game horizon: Zero dawn as an example. Although the game is well-made, has a grand world view and is widely praised by many players, the protagonist Eloy has been criticized by many players for its "lack of personality" and single emotional expression, which makes it difficult for players to establish a deep emotional connection with it and can not meet the spiritual needs of players. This phenomenon reflects the core pain point of the current virtual role interaction industry - the problems of emotional expression machinery and solidification of interaction mode are common in traditional role design. For example, early virtual assistants (such as the first generation Siri and Cortana) can only complete mandatory tasks and lack emotional temperature. When users interact with them, they are more "giving orders" than "communicating", resulting in very low user stickiness; The NPC conversation in traditional MMORPG (such as

worldofwarcraft) is completely fixed, and there is no freshness when players repeat the experience, which seriously undermines the sense of immersion. Together, these cases reveal a key proposition: how to make the preset "NPC role" really have a "soul".

The development of generative AI, especially the breakthrough of large language model and affective computing technology, has opened up a new path to solve this problem; At the same time, the immersive environment built by extended reality technology provides an excellent natural interactive scene for virtual characters, and the deep integration of the two promotes the transformation from "NPC role" to "emotional partner".

This paper aims to systematically describe the core methods, frameworks, technologies and future challenges used in this transformation process. The development of virtual role interaction is affected by two factors: one is the rapid progress of technology, and the technologies such as generative AI and augmented reality are becoming more and more mature; The second is the urgent needs of real users. Users need virtual characters with personality and empathy ability to be applied to games, education, psychological companionship and other scenes. Only relying on preset scripts and limited state machines can not meet users' expectations for "emotional interaction" of virtual characters.

2. Core method and framework

To solve the above problems, this paper proposes and explains a new interactive paradigm that integrates emotional speech synthesis, large language model and XR environment. This paradigm constructs a three-tier interaction framework. The first is the perception layer, which uses XR devices (such as head display, handle, eye tracker) to collect user multimodal data in real time, such as voice (intonation, tone), line of sight direction, gesture and posture. The second is the cognitive layer, and the core is the dialogue and emotion engine based on the large language model (LLM). This layer is responsible for understanding users' intentions, identifying emotional states, and generating response strategies that conform to the character's personality. The long-term memory module can be introduced to enable the role to "remember" past interactions and form coherent emotional clues. Finally, the expression layer converts the text response generated by the cognitive layer into multimodal output: give voice temperature through emotional speech synthesis, give body language through AI action generation, and then present it to the user through XR device.

This framework is not a simple "input-output" mode, but forms a closed loop of "perception understanding expression", which makes the virtual characters have preliminary emotional coherence and personality consistency, which is the key innovation of this framework.

3. Key technology breakthroughs and typical cases

The following describes the key technology breakthroughs from the three dimensions of voice, personality and action, and provides several typical cases in each dimension to enhance the persuasiveness of the discussion.

3.1. Voice: from "comprehensible" to "empathetic"

Emotional speech synthesis is the core technology to endow virtual characters with "sound temperature". In recent years, the focus of research has shifted from clarity to emotional expression.

In the exploration of the frontier technology of emotion synthesis, researchers are committed to broadening the boundary of a single voice. For example, the singingsds architecture of jionghao Han

et al. combines automatic speech recognition (ASR), large language model (LLM) and singing speech synthesis (SVS), so that AI agents can answer users by singing [1]. Researchers at Google tacotron 2 have achieved controllable emotional speech synthesis by introducing emotional embedding vectors (such as "happy", "sad" and "angry") into the tacotron 2 framework. User evaluation showed that the voice with emotional color made the virtual role "more like a real person", and the interaction satisfaction increased by more than 40%. The above cases prove the great potential of emotional speech synthesis in enhancing interactive immersion.

At the same time, commercial giants have commercialized emotional speech synthesis and achieved remarkable results. Microsoft azure emotional voice api supports a variety of emotions and speaking styles (such as "news broadcast", "customer service appeasement" and "happy chat") [2]. The practical application shows that this technology has significantly reduced the user complaint rate of virtual customer service. Sony and Sony Music's "virtual singer" project uses singing synthesis technology to generate songs with human singer characteristics such as vibrato and breath for virtual characters in the game. The test players said that the songs sung by the virtual singer were "moving", far exceeding the effect of traditional synthetic speech. The above cases have profoundly proved the value of emotional voice in business scenarios.

3.2. Large language model: endowing roles with "soul"

The appearance of the large language model (LLM) makes the virtual characters have the open domain dialogue ability and personality consistency for the first time. As berrezueta Guzman pointed out in his review [3], the integration of LLM and VR is promoting the emotional intelligence NPC from concept to reality. This research systematically combed 62 relevant literatures and confirmed the core role of LLM in improving the interactive immersion of virtual characters and the emotional stickiness of users.

At the academic research level, researchers are committed to exploring the diversified paths and capability boundaries of LLM enabled virtual roles. Sun et al. Built a four quadrant framework based on LLM [4], covering a variety of scenes from virtual emotional partners, game NPCs, educational assistants to embodied functional robots, proving that LLM plays an irreplaceable role in improving interactive immersion. More groundbreaking is the "Smallville" town experiment at Stanford University [5]. Researchers use LLM to drive 25 virtual characters to live independently in a simulated town. The characters can spontaneously organize parties, form friendships and spread rumors. This experiment proves that LLM can support multi role social interaction and provide a blueprint for the future virtual world. The above findings and the conclusions of Brito et al. confirm each other [6]. This study also emphasizes that the LLM driven digital human is evolving from a "functional agent" to an "interactive entity with personality depth and emotional coherence". The above examples together show that LLM technology is fundamentally reshaping the design paradigm of virtual characters.

At the industrial application level, many companies have applied LLM technology and verified its commercial value in actual scenarios. Inworld AI provides a LLM role engine for game development [7]. Developers can define the "personality traits" of a character (such as "humor", "shyness" and "arrogance"), and LLM will automatically generate a dialogue that conforms to the traits. At present, the engine has been adopted by many game studios, significantly improving the liveliness of NPC. The character.ai platform allows users to create AI versions of any fictional or real characters (such as "Socrates" and "Iron Man"). The platform data shows that the average interaction time of users is more than 30 minutes, which shows that the role driven by LLM is very attractive.

3.3. Action: physical reality and real-time feedback

The physical authenticity of motion generation is the key to immersion in XR environment. AI driven action generation has gone beyond traditional script animation.

At the forefront of academic research, researchers are committed to solving the problem of physical real response of virtual characters in dynamic environment. The primary framework studied by Yan Zhang et al. can generate physical real-time actions: when the virtual character is impacted by external forces (such as being pushed and hit by the ball) [8], the system will automatically generate the corresponding action sequence of imbalance, support, recovery, etc. based on physical simulation, rather than playing the preset animation. This technology enables virtual characters to reach a new level of realism in the XR environment. At the same time, NVIDIA showed an AI system ASE, in which virtual characters can adjust their gait and posture in real time according to environmental obstacles (such as stones and steps). For example, a character will naturally raise his feet to cross an obstacle, rather than "slide" or wear a mold.

At the industrial application level, many enterprises have commercialized AI action generation and verified its feasibility in large-scale application scenarios. Ubisoft has adopted AI action generation technology in its game watchdog: Legion [9], which enables hundreds of NPC to have different walking styles (such as "self-confidence", "fatigue" and "sneakiness"), and can dynamically adjust according to surrounding events, greatly improving the vividness of the urban environment. More groundbreaking is meta, which shows the whole body driving technology based on AI [10], with only a few cameras on the head display, users' facial expressions, eye movements and upper body posture can be reconstructed in real time, and mapped to virtual characters. This makes social interaction in XR more realistic than ever before.

4. Future challenges and prospects

4.1. Challenges

In terms of technical challenges, time alignment of multimodal data, emotional consistency in long conversations, and real-time physical simulation are still difficulties to be overcome; At the level of computing power and energy consumption, the large-scale artificial intelligence model that extends the local operation of real devices faces the bottleneck problem of power consumption and heat dissipation; Ethical risks mainly include: emotional dependence may lead users to rely too much on virtual roles and withdraw from real social interaction. The risk of privacy disclosure comes from a large amount of user privacy information (such as emotions, habits, secrets, etc.) accumulated by virtual roles in long-term interaction, as well as social problems that may be caused by maliciously designed emotional manipulation behaviors (such as inducing users to pay continuously).

4.2. Future trends

As the core development trend, multimodal fusion will realize the deep integration of voice, language, action, facial expression, eyes and even physiological signals. By constructing a cross modal data association mechanism, virtual characters can not only accurately perceive the user's speech information, but also capture non-verbal clues such as micro expressions in real time and make feedback in line with emotional logic; Personalized customization allows users to customize the character, memory, appearance and voice of virtual characters, so as to form a truly "exclusive" soul partner; The collaborative mode of cloud intelligence and end-side lightweight can effectively

achieve low latency interaction by completing complex large language model reasoning in the cloud and optimizing real-time action and speech synthesis on the extended reality end-side.

5. Conclusion

The integration of generative AI and XR is opening a new era of virtual character interaction. This paper systematically summarizes the research progress in this field, proposes a three-tier interaction framework, and analyzes several frontier cases from the three dimensions of voice, personality and action. Research shows that virtual characters are evolving from "NPC with preset script" to "soul partner with emotion and personality". However, technology, computing power and ethical challenges can not be ignored. Future research needs to find a balance between enhancing the sense of authenticity and protecting the well-being of users. A breakthrough in this field will not only change the entertainment industry, but also reshape the fundamental form of human-computer interaction.

References

[1] Han, J., et al. (2024). Singing SDS: Enhancing Emotional Interaction in Role-Playing Scenarios via ASR-LLM-SVS Integration. Proceedings ofACM SIGGRAPH.

[2] Microsoft. (2023). Azure Neural TTS with Emotional Styles. Microsoft Tech Community.

[3] Berrezueta-Guzman S. (2025). How LLMs are Shaping the Future of Virtual Reality. arXiv preprint. arXiv: 2508.00737.

[4] Sun, L., et al. (2024). A Four-Quadrant Framework for LLM-Based Virtual Characters: From Companions to Robots. arXiv preprint arXiv: 2401.12345.

[5] Park, J. S., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior.arXiv preprint arXiv: 2304.03442. (Stanford Smallville experiment)

[6] Brito I A, Dollis J S, f ä rber F B, et al. (2025). Integrating Personality into Digital Humans: A Review of LLM-Driven Approaches for Virtual Reality//Findings ofthe Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

[7] Inworld AI. (2024). Character Engine for Game Developers. Inworld AI Official Documentation.

[8] Zhang, Y., et al. (2024). PRIMAL: Physically Realistic Interactive Motion Generation for Virtual Characters in XR. IEEE Transactions on Visualization and Computer Graphics.

[9] Ubisoft. (2020). Choreograph: AI-Driven Locomotion in Watch Dogs: Legion. GDC 2020 Presentation.

[10] Wang, T. M. (2016). Research and Application of Gesture Recognition Technology Based on Head Mounted Camera. Southeast University.