A Latency-Optimized LLM-based Multimodal Dialogue System for Embodied Conversational Agents in VR

Konstantin W. Kühlem
RWTH Aachen University
Aachen, Germany
kuehlem@cs.rwth-aachen.de

Jonathan Ehret
RWTH Aachen University
Aachen, Germany
jehret@posteo.de

Torsten W. Kuhlen
RWTH Aachen University
Aachen, Germany
kuhlen@vr.rwth-aachen.de

Andrea Bönsch
RWTH Aachen University
Aachen, Germany
boensch@vr.rwth-aachen.de

Figure 1: Our LLM-driven guide assists users in exploring the museum and in learning about artworks and artists.

Abstract

Interactions with Embodied Conversational Agents (ECAs) are essential in many social Virtual Reality (VR) applications, highlighting the growing demand for free-flowing, context-aware conversations supported by low-latency, multimodal ECA responses. We introduce a modular, extensible framework powered by an Large Language Model (LLM), featuring streaming-based optimization techniques specially crafted for multimodal responses. Our system is capable of controlling self-behavior and task execution, in the form of moving through the Immersive Virtual Environment (IVE) directly controlled by the LLM, and is also capable of reacting to events in the IVE. In our study, our applied optimizations achieved a latency improvement of about 66% on average compared to having no optimizations.

ACM Reference Format:
Konstantin W. Kühlem, Jonathan Ehret, Torsten W. Kuhlen, and Andrea Bönsch. 2025. A Latency-Optimized LLM-based Multimodal Dialogue System for Embodied Conversational Agents in VR. In *ACM International Conference on Intelligent Virtual Agents (IVA '25), September 16–19, 2025, Berlin, Germany.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3717511.3749287

IVA '25, Berlin, Germany

ACM ISBN 979-8-4007-1508-2/25/09
https://doi.org/10.1145/3717511.3749287

1 Introduction

ECAs, present in many VR applications, are designed to mimic human behavior both verbally and non-verbally in face-to-face communication [6, 16]. They require seamless integration of components like Speech-to-Text (STT), decision-making modules, and Text-to-Speech (TTS) for dynamic and meaningful conversations. Traditionally, dialogue systems relied on pre-defined rule-based flows or decision trees [3], which result in limited conversational flexibility and demand significant manual effort due to their rigid structure. To overcome this, recent approaches leverage LLMs to generate human-like responses in real-time [5, 21, 23]. They have enabled use cases with agents in education [10, 14, 19], video games and game-like applications [8, 9, 20], and medicine [17]. Despite their capabilities, LLMs present technical challenges, particularly in latency, which is key for maintaining social presence and enhancing the user experience [4, 8]. There already exist mitigation attempts, including conversational fillers [4] and multimodal feedback [13]. Other approaches have crafted special components, like a TTS which operates on the tokens — the smallest unit of information — the LLM generates, exploiting the token-by-token generation nature of the LLMs [11, 12, 18]. In contrast to those attempts, our work focused on technically optimizing latency with given components by leveraging how they are interconnected in a complete pipeline design. Therefore, our approach uses streams the user's speech as well as the LLM output to achieve low-latency responses.

Multimodal abilities are also an important feature for the realism of ECAs [7]. While there exist LLM-based approaches that analyze the textual LLM response to select the behavior [15, 22], or utilize

multiple LLMs [1], we implemented JSON-structured responses to control multimodal features of the ECA through a single LLM response, allowing the LLM to fully control the multimodal aspects, which in our case are self-behavior and moving through the IVE.

Our LLM-driven solution is applicable to any ECA. For this demo, we chose a virtual museum guide use case (e.g., [2]) due to its need for interactive assistance and appropriate answers to potentially complex or unplannable inquiries.

2 System

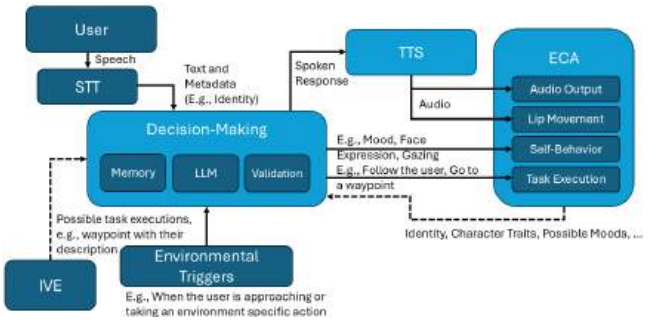

Figure 2: Architecture of our LLM-driven dialogue system, detailing component integration and data flow.

Our dialogue system, seen in Fig. 2, is implemented in a modular fashion and centered around the general pipeline of input, decision-making ('thinking'), and output. The input to the system can either originate from the user's speech or an environmental trigger. Speech input is processed by an STT component, located in the top left of Fig. 2, transcribing spoken utterances into text while the user is talking. We tested several STT services and chose the Google Cloud STT based on quality, latency, and the fact that it enables us to stream the audio data in real time. In our user study, with $N = 30$ participants, the STT service took on average 0.350 ± 0.157 s ($M \pm SD$) in 599 speech-triggered pipeline executions, which in our tests, using texts from our user study, proved more than five times faster than without streaming. Environmental triggers — non-verbal stimuli — enable the ECA to respond to specific events within the IVE, like the user approaching, reaching a designated area, or other context-dependent scenarios.

The decision-making component, located centrally on the left side of Fig. 2, is powered by an LLM, allowing for dynamic and unscripted conversations. For the LLM, we chose GPT-4o, as it was the newest GPT model available at the time, and many related works rely on this family of LLMs. In our tests, this model also provided a better perceived latency quality and better handling of the multimodal behavior. To support natural user-ECA interactions, a short-term memory mechanism is integrated, enabling the system to "remember" previous parts of the conversation. The responses by the LLM may consist of a response text, self-behavior, and task execution — together forming the multimodal behavior — which is achieved by instructing the LLM to answer in the JSON format. In our work, self-behavior is expressed through facial expressions — reflecting the mood of the ECA — as well as through gestures. Task execution enables the ECA to navigate the IVE, either by moving to specific locations or by following the user. To achieve this, the LLM can use five possible JSON fields: *response*, *face*, *gesture*, *move_to*, and *attention*. For the fields *face* and *gesture*, the LLM can choose from nine predefined facial expressions and two predefined gestures. The *attention* field allows the ECA to share its visual attention between the user and a point of interest, like a painting, if necessary. The response of the LLM is streamed, which is possible due to the nature of the LLM generating token by token, as explained previously. With a specially crafted JSON parser, the mentioned JSON fields are available to the next component as soon as they are generated. Additionally, the spoken response is sent to the TTS component on a sentence-by-sentence basis, deviating from a purely sequential pipeline design. This approach reduced processing time for the LLM component by 0.461 ± 0.339 s on average, measured in our user study in 882 pipeline executions (speech and environmental triggered). For the TTS, we chose Amazon Polly for this task due to its low latency, which was measured at 0.272 ± 0.337 s in our study. The audio is also used to drive the real-time lip-syncing behavior.

With both optimization techniques, the overall response latency is drastically decreased by around 66% compared to no optimizations. In absolute terms, the system delivered a response in about 1.554 ± 0.668 s on average in our user study in the 599 speech-triggered executions.

3 Demonstration

Our demo features an LLM-driven MetaHuman guide within a virtual museum (see Fig. 1), allowing users to engage in open-ended conversations while exploring artworks. Users navigate the museum using a Quest 3 and two controllers, with the primary controller for hand-directed steering and the secondary controller displaying a to-do list, informing about potential tasks and goals in the museum. The museum features well-known real-world paintings, presented in four thematically organized halls such as Surrealism, Self-Portraits, and Landscape Painting. Users can converse freely with the guide about the artworks, artists, and hall themes, and may be guided directly to points of interest. The environment includes assets from the Museum Environment Kit[1], the National Maritime Museum[2] package, and the City Park Environment Collection[3].

Conversations occur through natural speech, enhancing presence and fluid communication with the virtual guide. Thereby, users experience variations in the guide's response latency via our two streaming optimization techniques. A button on the secondary controller enables users to toggle settings for STT and LLM processing, cycling between enabling both streaming optimizations, activating only one, or disabling them entirely. This promotes real-time experimentation with conversational dynamics, highlighting the impact of these optimizations on naturalness and flow of the interaction. To reflect this to the user, the to-do list also provides information on technical metrics, such as the latency of the last response and how the LLM used the multimodal fields.

[1] https://www.unrealengine.com/marketplace/museum-environment-kit
[2] https://www.unrealengine.com/marketplace/national-maritime-museum
[3] https://www.unrealengine.com/marketplace/city-park-environment-collection

References

[1] Andrea Bellucci, Giulio Jacucci, Kien Duong Trung, Pritom Kumar Das, Sergei Viktorovich Smirnov, Imtiaj Ahmed, and Jean-Luc Lugrin. 2025. Immersive Tailoring of Embodied Agents Using Large Language Models. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE. doi:10.1109/VR59515.2025.00063

[2] Andrea Bönsch, David Hashem, Jonathan Ehret, and Torsten W. Kuhlen. 2021. Being Guided or Having Exploratory Freedom: User Preferences of a Virtual Agent's Behavior in a Museum. In *Proc. of the 21th ACM International Conference on Intelligent Virtual Agents*. doi:10.1145/3472306.3478339

[3] Tibor Bosse and Simon Provoost. 2015. Integrating conversation trees and cognitive models within an eca for aggression de-escalation training. In *PRIMA 2015: Principles and Practice of Multi-Agent Systems: 18th International Conference, Bertinoro, Italy, October 26-30, 2015, Proceedings 13*. Springer, 650–659.

[4] Halim-Antoine Boukaram, Micheline Ziadee, and Majd F Sakr. 2021. Mitigating the Effects of Delayed Virtual Agent Response Time Using Conversational Fillers. In *Proceedings of the 9th International Conference on Human-Agent Interaction*. doi:10.1145/3472307.3484181

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020). doi:10.48550/arXiv.2005.14165

[6] Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001). doi:10.1609/aimag.v22i4.1593

[7] Justine Cassell, David McNeill, and Karl-Erik McCullough. 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition* 7, 1 (1999). doi:10.1075/pc.7.1.03cas

[8] Frederik Roland Christiansen, Linus Nørgaard Hollensberg, Niko Bach Jensen, Kristian Julsgaard, Kristian Nyborg Jespersen, and Ivan Nikolov. 2024. Exploring Presence in Interactions with LLM-Driven NPCs: A Comparative Study of Speech Recognition and Dialogue Options. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*. Association for Computing Machinery. doi:10.1145/3641825.3687716

[9] Samuel Rhys Cox and Wei Tsang Ooi. 2023. Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback. In *International Workshop on Chatbot Research and Design*. Springer, Springer Nature Switzerland. doi:10.1007/978-3-031-54975-5_10

[10] Laduona Dai, Merel M. Jung, Marie Šafář Postma, Janneke van der Loo, and Max M. Louwerse. 2024. LittleGenius: Co-Designing a GPT-4 Enhanced VR Pedagogical Framework with Teachers for Primary Education. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '24)*. ACM. doi:10.1145/3652988.3673941

[11] Trung Dang, David Aponte, Dung Tran, Tianyi Chen, and Kazuhito Koishida. 2025. Zero-Shot Text-to-Speech from Continuous Text Streams. https://openreview.net/forum?id=RK3Gj9J5my

[12] Avihu Dekel, Slava Shechtman, Raul Fernandez, David Haws, Zvi Kons, and Ron Hoory. 2023. Speak While You Think: Streaming Speech Synthesis During Text Generation. doi:10.48550/arXiv.2309.11210

[13] Morad Elfleet and Mathieu Chollet. 2024. Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM Powered Embodied Conversational Agents in Virtual Reality. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '24)*. ACM. doi:10.1145/3652988.3673965

[14] Marco Fiore, Michele Guattullo, and Marina Mongiello. 2024. First Steps in Constructing an AI-Powered Digital Twin Teacher: Harnessing Large Language Models in a Metaverse Classroom. *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (Mar 2024). doi:10.1109/VRW62533.2024.00266

[15] Hanseob Kim, Ghazanfar Ali, Seungwon Kim, Gerard J. Kim, and Jae-In Hwang. 2021. Auto-generating Virtual Human Behavior by Understanding User Contexts. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 591–592. doi:10.1109/vrw52623.2021.00178

[16] Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. 2017. Embodied Conversational Agents in Clinical Psychology: A Scoping Review. *J Med Internet Res* 19, 5 (09 May 2017). doi:10.2196/jmir.6553

[17] Alexandra Retevoi, Giada Devittori, Tobias Kowatsch, and Olivier Lambercy. 2024. Building Conversational Agents for Stroke Rehabilitation: An Evaluation of Large Language Models and Retrieval Augmented Generation. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '24)*. ACM. doi:10.1145/3652988.3673924

[18] Sambal Shikhar, Mohammed Irfan Kurpath, Sahal Shaji Mullappilly, Jean Lahoud, Fahad Khan, Rao Muhammad Anwer, Salman Khan, and Hisham Cholakkal. 2025. LLMVoX: Autoregressive Streaming Text-to-Speech Model for Any LLM. doi:10.48550/arXiv.2503.04724

[19] Mae Taylor, Zaid Muwaffak, Matthew Penny, Blanka Szulc, Steven Brown, Andy Merritt, and Stephen Hilton. 2023. The Rise of the AI Scientist: Unleashing the Potential of Chat-GPT Powered Avatars in Virtual Reality Digital-twin Laboratories. (Nov 2023). doi:10.26434/chemrxiv-2023-t4vg7

[20] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building LLM-based AI Agents in Social Virtual Reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–7. doi:10.1145/3613905.3651026

[21] Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023. An Survey of the Evolution of Language Model-Based Dialogue Systems. *arXiv preprint arXiv:2311.16789* (Nov 2023). doi:10.48550/arXiv.2311.16789

[22] Takato Yamazaki, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiki Kawamoto, and Toshinori Sato. 2023. An Open-Domain Avatar Chatbot by Exploiting a Large Language Model. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 428–432. doi:10.18653/v1/2023.sigdial-1.40

[23] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv:2303.18223* (2023). doi:10.48550/arXiv.2303.18223