ACM Symposium on Applied Perception 2026

Phase-Based Latency Mitigation in LLM-Driven VR Agents

Konstantin W. Kühlem, Ying Zhou, Torsten W. Kuhlen, Andrea Bönsch
Visual Computing Institute, RWTH Aachen University

Introduction

Background

- Response latency in LLM-driven Virtual Agents (VAs) breaks conversational flow, making the VA seem inattentive, lowering perceived naturalness, intelligence, & rapport [Boukaram2021, Elfleet2024]
- Approaches to reduce response latency
  → *Technical optimizations* of LLM pipelines reducing the actual processing time (e.g., [Kühlem2026])
  → *Behavioral Latency Mitigation (BLM)* with multimodal VA cues targeting how waiting is perceived (e.g., [Elfleet2024])

Research Gap

- Existing work treats BLM as one waiting block, not exploiting separate cues for the User-Speaking Phase (USP) and the Response-Waiting Phase (RWP)

Research Objective

- Assessing perceived latency and social-quality metrics to determine which phase (USP vs. RWP) provides the strongest mitigation effects in VR

Setting

Dual LLM Architecture

- Full-size LLM (GPT-4.1): processes user utterance chunks to generate VA answer [replicating Kühlem2026] - latency fixed to 4 sec
- Fast LLM (GPT-4.1 nano): runs concurrently, streams short filler cues (back-channel vocalizations, “thinking” prompts) every 0.6 sec providing USP/RWP multimodal cues (2 sec cooldown for vocalization; non-verbal cues limited by Fast LLM capacity)

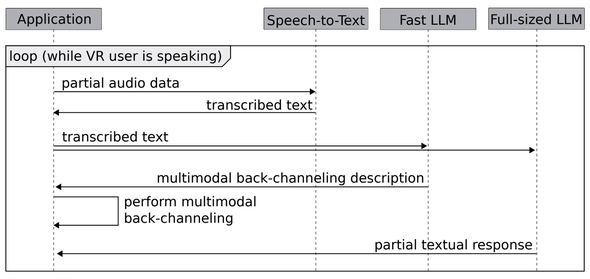


Fig.1: USP - user speaking phase, while VA is actively listening

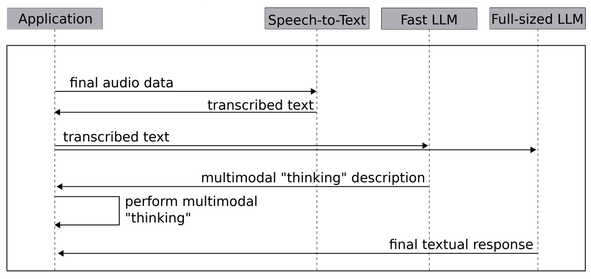


Fig.2: RWP - response-waiting phase: VA is preparing to answer, while VR user waits

USP Cues

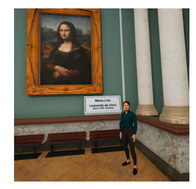
- Directed gaze toward the user, brief nods, and subtle smile, light hand-lifts & low-amplitude gestures, back-channeling (e.g., “uh-huh”) at a lower volume than the final answer

RWP Cues

- “Thinking” posture with slow head tilt, hand-to-chin, and slight gaze shifts, short verbal cue (“Let me think.”, ...) louder than USP cues but below the final response volume

Immersive Virtual Environment [Kühlem2026]

- UE5.6-based indoor museum with curator Vivian
- Meta Quest 3 experience



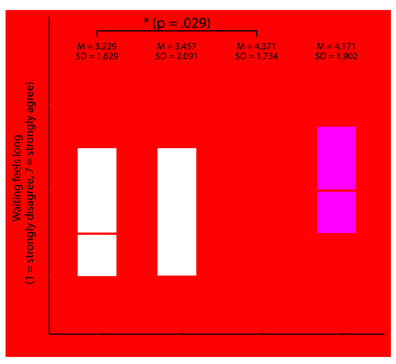
VR-based Evaluation

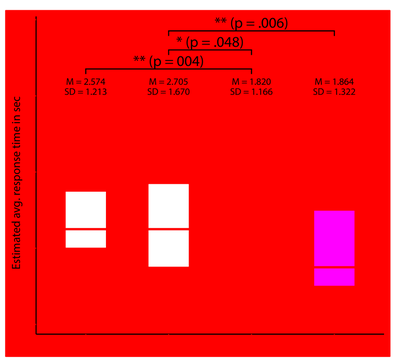
Design

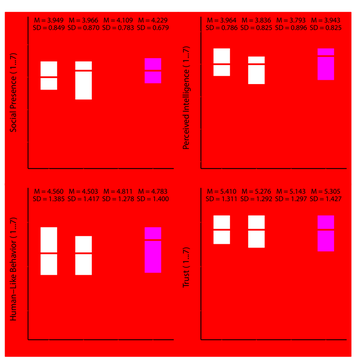
- Within-subject with four randomized conditions
  C_{no} (no feedback) C_U (USP-only) C_R (RWP-only) C_B (USP + RWP)
- Free conversation with VA Vivian for 5 min per condition
- 35 participants (20M/15F, 26.29 ± 3,46 years)
- Per-condition subjective measurements:
  - 7-pt Likert item “Waiting for Vivian’s response feels long.”
  - Numeric estimate of avg. response time in sec
  - Evaluation of interaction quality via
    [Social Presence], [Perceived Intelligence],
    and [ASAQ]’s human-likeness & trust constructs
- Post-study subjective measurements:
  Open-ended questions on pros/cons of BLM techniques

Key Quantitative Findings

- C_R strongest reduction of perceived waiting (sign. higher than C_{no}) with C_U & C_B: modest, non-significant improvements over C_{no}
- Estimated response latency was lowest in C_R and C_B while C_R was significantly lower than C_{no} and C_U and C_B was significantly lower than C_U
- Social quality scores revealed no significant effects







Qualitative Insights

- “Thinking” gestures (C_R, C_B) perceived as helpful for a more natural and understandable waiting period (~42% of comments)
- C_U feedback described as distracting and interfering with conversation flow (~18% of comments)

Take-Aways

- Feedback during RWP cuts perceived latency the most
- Adding cues during USP offers no extra benefit and can even be distracting

Literature

Boukaram et al., 2021
Mitigating the Effects of Delayed Virtual Agent Response Time Using Conversational Fillers.
In Int. Conf. Human-Agent Interaction. doi:10.1145/3472307.3484181

Elfleet and Chollet, 2024
Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. *In ACM Int. Conf. Intell. Virtual Agents*, doi:10.1145/3652988.3673965

Kühlem et al., 2026
Fostering Engagement Through a Latency-Optimized LLM-Based Dialogue System for Multimodal ECA Responses.
In IEEE Int. Conf. on Artif. Intell. eXtended Virtual Reality. doi:10.1109/AIxVR67263.2026.00019

Perceived Intelligence
Bartneck et al., 2009: Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *In Int. J. Soc. Robotics*. doi:10.1007/s12369-008-0001-3

ASAQ
Fitrianie et al., 2022: The Artificial-Social-Agent Questionnaire: Establishing the Long and Short Questionnaire Versions.
I*n ACM Int. Conf. Intell. Virtual Agents*. doi:10.1145/3514197.3549612

Social Presence
Makransky et al.. 2017: Development and Validation of the Multimodal Presence Scale for Virtual Reality Environments: A Confirmatory Factor Analysis and Item Response Theory Approach. *In Comput. Hum. Behav*. doi:10.1016/j.chb.2017.02.066

Visual Computing Institute
Virtual Reality & Immersive Visualization
Prof. Dr. Torsten W. Kuhlen

Funded by the DFG within the project "Examining Attention, Memory Performance, and Listening Effort (exAMPLE)" (444724862), which is part of the DFG Priority Program
AUDICTIVE SPP2236

RWTH AACHEN UNIVERSITY