POLITECNICO DI TORINO
Repository ISTITUZIONALE

Investigating player perception of environment-aware embodied conversational agents enabled by vision language models

Original
Investigating player perception of environment-aware embodied conversational agents enabled by vision language models / Fiorenza, J., Thawonmas, R., Calandra, D., Lamberti, F.. - ELETTRONICO. - (2026), pp. 1-6. (2026 IEEE 5th International Conference on Intelligent Reality (ICIR 2026) Pisa (IT) June 25 - 26, 2026) [10.1109/ICIR70631.2026.11628135].

Availability:
This version is available at: 11583/3009647 since: 2026-08-15T16:22:31Z

Publisher:
IEEE

Published
DOI:10.1109/ICIR70631.2026.11628135

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

19 November 2026

Investigating Player Perception of Environment-Aware Embodied Conversational Agents Enabled by Vision Language Models*

Jacopo Fiorenza, Ruck Thawonmas, *Senior Member, IEEE*,
Davide Calandra, *Member, IEEE*, and Fabrizio Lamberti, *Senior Member, IEEE*

***Abstract*—Recent advancements in generative AI have enabled the integration of Large Language Models (LLMs) into Virtual Environments (VEs). This integration has been particularly useful to create Embodied Conversational Agents (ECAs) capable of engaging in meaningful interactions. However, such ECAs often lack environmental awareness, which may limit the quality of interactions and conversations. A potential solution could be to use Vision Language Models (VLMs), which could empower ECAs with structured environmental knowledge. However, the use of VLMs for such applications is largely underexplored. This paper explores how VLMs can be employed to integrate environmental awareness in ECAs and investigates their impact on player perception. To this end, an architecture leveraging multi-image inference and scene graph generation was designed. A within-subject user study was then conducted in a virtual reality game, comparing ECAs with and without environment-aware capabilities. The evaluation considered several dimensions, including perceived knowledge, intelligence, factuality, willingness of future interaction and conversational abilities.**

***Index Terms*—Games, Virtual Reality, Conversational Agents, LLMs, VLMs**

I. Introduction

Recent advances in generative AI have enabled the integration of Large Language Models (LLMs) into Virtual Environments (VEs) [1], fostering the development of Embodied Conversational Agents (ECAs) capable of engaging in natural interactions.

Given their capabilities, LLMs have been increasingly adopted to create conversational Non-Player Characters (NPCs) in games and VR applications, potentially enhancing immersion through human-like dialogue [2], [3].

However, most LLM-based ECAs lack direct awareness of their surrounding VE. In traditional gaming applications, environmental knowledge is typically injected through manually defined textual descriptions or structured representations extracted from the game engine [4]–[6]. While effective, these approaches require explicit modeling of objects and spatial relations, which may become costly in complex VEs [5].

A potential alternative is to employ Vision Language Models (VLMs), which jointly process visual and textual inputs. By extracting structured representations directly from visual frames, VLMs could enable automatic environment perception without manual scene annotation. Despite their growing capabilities, their use for empowering LLM-based ECAs with environmental awareness in gaming contexts remains underexplored, particularly with respect to player perception. However, empirical evidence on how visually grounded ECAs influence user perception remains limited.

This paper investigates whether integrating a VLM-based Environment Perception Module (EPM) into an LLM-powered ECA affects player perception in a virtual game scenario. The proposed module performs multi-image inference to generate a scene graph representation of the environment, which is injected into the conversational system prompt of the agent.

The approach was evaluated through a within-subject user study comparing ECAs with and without environment-aware capabilities. The analysis considers perceived knowledge, intelligence, factuality, willingness of future interaction, and conversational qualities.

The objectives of this study are summarized in the following research questions:

- `R1`: How environment knowledge enabled by a VLM impacts players' perception of an LLM-ECA in terms of perceived knowledge and specific perceived knowledge of the environment?
- `R2`: How environment knowledge enabled by a VLM impacts players' perception of an LLM-ECA in terms of perceived intelligence, environment factuality and willingness of future interaction?
- `R3`: How environment knowledge enabled by a VLM impacts players' perception of an LLM-ECA in terms of conversational capabilities, such as communicative ability, flexibility and human-likeness?

Overall, the contribution lies in a lightweight VLM-based module that derives scene knowledge from visual frames as an alternative to manually authored or engine-dependent state representations, and in an exploratory evaluation of its perceptual impact in an interactive game scenario. The focus of this work is therefore on perceived environmental grounding and interaction quality, rather than on objective validation against engine-level ground truth.

*This publication is part of the project PNRR-NGEU which has received funding from the MUR - DM. 352/2022.

J. Fiorenza, D. Calandra, and F. Lamberti are with the Politecnico di Torino, Italy (e-mail: {name}.{surname}@polito.it).
R. Thawonmas is with College of Information Science and Engineering, Ritsumeikan University, Japan (e-mail: ruck@is.ritsumei.ac.jp)
(Corresponding author: Jacopo Fiorenza.)

II. Related Works

Conversely, most HAI studies involve direct interaction between LLM-based agents and users. For instance, Pan et al. [7] developed ELLMA-T, an ECA specifically meant for language learning in social VR. The ECA was able to engage in meaningful conversations with the user utilizing an LLM (GPT-4), Speech-To-Text (STT) to process input data and Text-To-Speech (TTS) to convert responses into audible feedback. Christiansen et al. [2] integrated several ECAs as conversational NPCs in a VR video game. The ECAs were designed to role-play while interacting with players, engaging in immersive conversations by using speech or dialogue options. Similarly, Yang et al. [3] proposed a similar ECA architecture by employing prompt engineering techniques, aiming to explore the impact on user's perception of ECAs with different level of knowledge.

Despite the advancements, current trends in HAI highlight an increasing need of integrating capabilities in the underlying architectures to improve not only ECAs knowledge and interaction skills, but also their context-awareness.

The architecture of ECAs relies on the presence of some key modules implementing specific capabilities. These modules are well highlighted in the work of Xi et al. [8], which defined the core of LLM-agents as consisting of three main modules: perception, brain, and action. Particularly, the perception module allows agents to perceive their surrounding stimuli, such as auditory and visual cues.

When it comes to ECAs in VEs, several approaches to allow this kind of environmental-knowledge have been explored. An approach is to feed the agent with specific game information through in-game events. This is typically made possible through elements that dynamically prompt the agent with contextual data [9]. An example is given by Z. Li et al. [4]. By employing an LLM (GPT4), the work showed how structured prompts can provide reliable environmental information to the ECA within the VE. A similar approach was adopted by H. Li et al. [5], who integrated a world state module in their ECA prompt engineering structure. This module was dynamically updated using a SG representation of the environment extracted from the game system.

Considering the above works, textual information appears to be the common way to make ECAs environment-aware, although some visual features (e.g., shapes, materials) must still be predetermined. In this context, the proposed approach is not intended to replace engine-level representations in terms of absolute accuracy, but to assess whether a scalable VLM-based pipeline can influence user perception.

Since current LLMs have the capability to process multimodal information, a viable solution could be to employ VLMs. This type of vision models has already proven to be effective in several scenarios involving a VE. An example is provided by Hebri et al. [10], who introduced SmartSage, an ECA for cognitive assessment in immersive VR. The system relied not only on LLMs for natural language interaction but, crucially, on VLMs to interpret contextual cues within the VE. Similarly, the work of Konenkov et al. [11] proposed a VLM-based agent for a medical training scenario. Through a fine-tuning approach, the VLM was employed to gather environmental information by inferring from images taken from different points of view of the VE.

Despite the advantages of employing VLMs within VEs, this kind of application field is still underexplored. This gap motivates the present investigation into perception effects of visually grounded ECAs.

III. System Design

This section reports the design of the proposed approach to empower virtual ECAs with environmental perception. Particularly, it highlights the integration of an Environment Perception Module (EPM) into an LLM-ECA architecture, which involves a VLM extracting a SG representation of the 3D environment through multi-image inference. Additionally, it also details the prompt-engineering modalities to obtain such SG representation.

A. System Architecture

As highlighted from the state of the art, when it comes to constructing conversational LLM-ECAs, traditional architectures involve several modules implementing different capabilities [12], [13]. To enable players' interaction with the system, usual practices encompass Speech Recognition (SR) and Speech-to-Text (STT) to translate conversations into text [14]. This information is then transmitted as a user prompt to a conversational module, typically involving an LLM, a system prompt containing textual instructions and information, and a memory buffer to keep track of the conversations. The LLM textual response is finally translated into audio through a Text-to-Speech (TTS) module, allowing the ECA to provide a response to the player.

To empower LLM-ECAs with environment-aware capabilities, the proposed approach integrates an EPM into their architecture. As depicted in Figure 1, the pipeline starts with the system within the platform (e.g., a game engine) performing the extraction of visual data from the VE. By taking advantage of one or more Virtual Cameras (VCs) placed into the 3D-VE, it asynchronously extracts N-frames representing different angles of a virtual scene. Once this step is completed, the frames are passed to a VLM module equipped with a system prompt to generate a SG representation of the VE under the form of a JSON scheme. Finally, the system prompt of the conversational module is updated with this SG.

The resulting architecture allows the conversational module to provide contextual answers to the players, thus possibly demonstrating to be aware of its surrounding VE. Depending on the static or dynamic nature of the VE (i.e., whether it presents changes over time or not) and on the response speed of the selected VLM, the SG generation process and the system prompt update may be repeated on the fly at any interval of time.

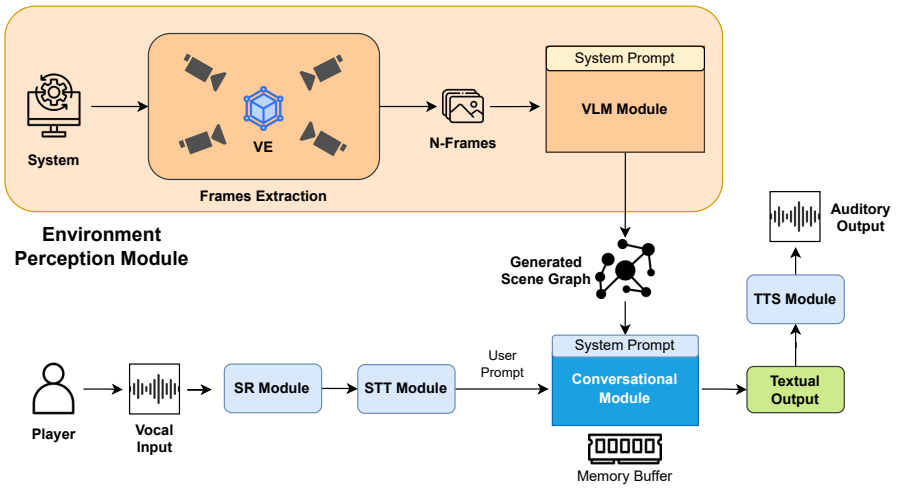


Fig. 1: Devised VLM-based environment perception module integrated in a traditional architecture of an LLM-based ECA. Over both the VLM module and the Conversational Module, their respective system prompts can be observed.

B. Prompts Design

Taking inspiration from state-of-the-art models [6], [15], [16], the task assigned to the VLM module is to generate a scene graph (SG) describing all visible objects in the VE from the extracted frames.

The system prompt specifies the task context, the multi-image input assumption, and a structured JSON output describing objects, attributes, and spatial relations. The design follows prior scene graph generation approaches while enforcing constraints on naming consistency and predicate sets to ensure structured and controllable outputs.

Based on these instructions, the VLM produces a structured representation of the environment.

For each detected object, the VLM generates a node containing its name and a list of attributes. The attributes describe relevant visual properties (e.g., color and material). The attribute schema is inspired by open-vocabulary attribute detection approaches such as Open-vocabulary Attribute Detection (OvAD) [15].

The module also generates edges representing spatial relations in subject-predicate-object form, limited to proximity and containment relations (e.g., near, on, under, inside, hanging from) [16].

The resulting SG is injected into the conversational module system prompt, enabling the LLM-based ECA to incorporate environmental information during interaction.

IV. Use Case

To validate the proposed approach, we integrated the Environment Perception Module (EPM) into an open-source murder-mystery game featuring LLM-powered NPCs [2]. The application was deployed in a desktop VR setup.

The proposed EPM was integrated within the architecture of the NPCs present in the video game, thus endowing them with VLM-powered environmental knowledge.

Figure 2 provides a concrete example of the proposed pipeline, showing a captured frame from the virtual environment, an environment-related interaction with the NPC, and an example of the generated scene graph used to produce a context-aware response.

A. Game Application

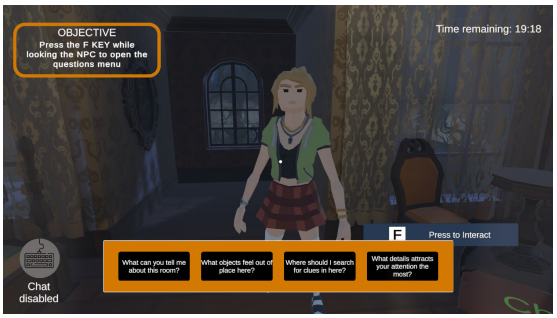
Within the game, players interact with multiple NPCs to gather information and solve a homicide. The VE is set in a mansion composed of multiple rooms populated with furniture and interactive objects.

The ECAs are represented by NPCs placed in different rooms of the VE, each one empowered with conversational capabilities based on LLMs (GPT4), STT, and TTS (Wit.ai) tools. Following prompt-engineering design principles, the authors [2] created a structured system prompt, communicating to the conversational module of each NPC the setup of the game (e.g., plot, location), their personality, and their alibi for the homicide. According to an event-based mechanism, the system prompt of each NPC was updated when the player found certain interactive objects, or clues, within the VE. Additionally, the system prompt included a set of rules that the NPC had to follow to allow a smooth conversation with the players (e.g., "*Do not mention that you are an NPC*"). Besides the above initialization information, the architecture of the NPCs was not given any form of environmental knowledge beforehand.

By employing the same Unity version as the original game (Unity 2022.3.9f1), the proposed EPM was embedded within the architecture of the NPCs. Since the NPCs occupied one room each, four VCs at different angles were placed in every one of them. The position and orientation of the cameras were defined to capture as much as possible the elements and visual features of the VE. Each camera is able to capture four frames representing different angles of the room they are in.



(a) Captured frame from the VE



(b) Environment-related interaction

```
{"objects":[
 {"name":"desk","attributes":["material: wood"]},
 {"name":"drawer","attributes":["state: closed"]}
],
"relations":[
 {"subject":"drawer","predicate":"inside","object":"desk"}
]}

NPC: “You could hide it inside the drawer under the desk.”
```

(c) Structured scene knowledge and dialogue use

Fig. 2: Overview of the proposed environment-awareness pipeline. (a) A frame captured from the VE and provided to the VLM module. (b) An example of environment-related interaction between the player and the NPC. (c) Excerpt of the generated scene graph and a corresponding dialogue response showing how structured environmental knowledge is incorporated into the NPC’s answer.

For each NPC room, multiple virtual cameras captured frames from different viewpoints at initialization.

Since the VE and the NPCs are static in the considered game, i.e., no spatial changes occur throughout the experience, frames are captured once at game initialization. The frames are then encoded into JPG format and sent to the VLM module for generating the SG in JSON format. For this purpose, the implementation developed in this study employs GPT-4.1 as VLM for the EPM, both for processing the frames and generating the SG. The SG is then forwarded to the NPC’s conversational module by appending a dedicated textual rule to its system prompt, providing instructions on how to integrate environmental information during interaction.

Captured frames were processed by a VLM (GPT-4.1) to generate a JSON-based scene graph, which was appended to the NPC system prompt to enable environment-aware responses. A single VLM/LLM configuration was adopted to isolate the effect of environmental awareness from model variability, as the goal of this study is not model benchmarking but perception analysis.

Additional adjustments were made to enhance interaction and accessibility within a controlled experimental setup. The evaluation was conducted in a single desktop VR configuration to reduce device-related variability. GPT-4.1 was used for dialogue, Whisper API for STT, and ElevenLabs Flash V2.5 for TTS due to faster responses than Wit.ai. Besides the original voice-based modality, a chat-based interface was introduced, allowing players to type questions if preferred to voice-interaction.

V. Experimental Procedure

The evaluation consisted of a within-subject study comparing NPCs without (Version A) and with (Version B) environmental knowledge. The order of the two conditions was counterbalanced across participants to mitigate order effects.

The study involved 15 participants with prior experience in desktop VR gaming. Given the sample size, the study is intended as exploratory, with a within-subject design chosen to increase sensitivity while limiting inter-participant variability. Each participant interacted with both versions and completed questionnaires after each condition.

Based on the research questions of the study, items for Perceived Knowledge, Specific Perceived Knowledge, Perceived Intelligence, Factuality, and Willingness of Future Interaction were adapted from a previous work [3]. To evaluate the conversational capabilities of the NPCs, the Partner Modeling Questionnaire (PMQ) [17] was employed. The two questionnaires contained respectively 17 and 18 questions on a 7-point Likert scale, for a total of 35 questions. After the questionnaires, the participants were asked to re-play the game with the other version of the NPCs (B or A), and re-compile the questionnaires ($\sim$20+10 minutes). The total time for each session was about $\sim$1 hour and 10 minutes. However, the relatively small sample size limits statistical power, particularly across multiple perceptual constructs.

VI. Results and Discussion

The analysis was carried out across seven dimensions derived from the employed questionnaires. The first ques-

tionnaire was used to evaluate Perceived Knowledge (PK), Specific Perceived Knowledge (SPK), Perceived Intelligence (PI), Factuality (FCTY), and Willingness for Future Interaction (WFI). The second questionnaire (PMQ) was analyzed to assess Communicative Competence and Dependability (CCD), Human-Likeness in Communication (HLC), and Communicative Flexibility (CF). To compare Version A and Version B, average participant scores were calculated for each dimension with respect to the corresponding version. The analysis was performed considering both the individual questions and the overall values for each dimension. The Shapiro-Wilk test was applied to assess the normality of each dimension. Depending on the outcome, either a Paired Samples t-test (for normally distributed data) or a Wilcoxon Signed-Rank test (for non-normal data) was conducted. In both cases, a 5% significance threshold ($p < .050$) was adopted to detect statistical differences between the two versions.

A. Version A vs Version B: First Questionnaire

The analysis of results for the first questionnaire indicates consistent differences between the two versions. In the overall comparison (Figure 3), Version B, which contained the NPCs with environmental knowledge, received higher ratings than Version A across all evaluated dimensions. The comparison of mean scores clearly shows higher values for Version B (M_{PK} = 5.67, σ_{PK} = .88; M_{SPK} = 6.18, σ_{SPK} = .61; M_{PI} = 6.13, σ_{PI} = .59; M_{FCTY} = 5.90, σ_{FCTY} = .83; M_{WFI} = 6.27, σ_{WFI} = .96) in comparison to Version A (M_{PK} = 3.92, σ_{PK} = 1.67; M_{SPK} = 4.05, σ_{SPK} = 1.59; M_{PI} = 5.37, σ_{PI} = .79; M_{FCTY} = 3.87, σ_{FCTY} = 1.81; M_{WFI} = 5.33, σ_{WFI} = 1.72), with statistical differences for PK ($p = .008$), SPK ($p < .001$), PI ($p = .005$) and FCTY ($p = .003$). Conversely, no statistical differences were found for the WFI dimension.

Overall, to answer the first two research questions (R1, R2), these results indicate that integrating environmental knowledge substantially improved participants' perception of the NPCs knowledge, intelligence and factuality, while willingness for future interaction did not show statistically significant differences. This may also be influenced by the limited statistical power for detecting smaller effects.

B. Version A vs Version B: Second Questionnaire (PMQ)

For the second questionnaire (PMQ), the overall comparison of mean scores (Figure 4) partially favors Version B, but the significance varies across dimensions. The aggregated results for CCD shows the clearest differences: Version B was consistently rated higher (M_{CCD} = 5.92, σ_{CCD} = .62) than Version A (M_{CCD} = 4.81, σ_{CCD} = 1.32) with a statistical difference ($p = .006$). In contrast, no statistical differences were found for HLC and CF between Version A (M_{HLC} = 4.87, σ_{HLC} = 1.27; M_{CF} = 4.84, σ_{CF} = 1.23) and Version B (M_{HLC} = 4.71, σ_{HLC} = 1.14; M_{CF} = 5.00, σ_{CF} = 0.91).

To address the third research question (R3), the findings suggest that integrating environmental knowledge mainly enhanced players' perceptions of competence and reliability, whereas human-likeness and communicative flexibility were largely unchanged. This result may be explained by the fact that perceptual dimensions such as HLC and CF do not depend solely on the NPC's knowledge of the environment, but also on its interaction and conversational modalities (e.g., human-like body movements, eye gaze, and speaking style). Furthermore, the strong generative capabilities of the LLM used in both conditions (GPT-4.1) may have reduced the likelihood of observing significant differences between them. Accordingly, non-significant differences in HLC and CF should be interpreted with caution, as smaller effects may not have been detectable under the present sample size and experimental conditions.

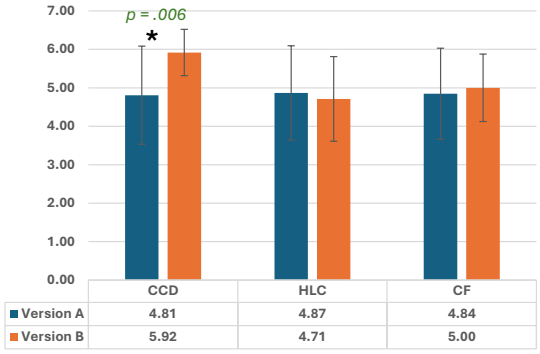


Fig. 3: Comparison of Version A and Version B across dimensions of the first questionnaire: Perceived Knowledge (PK), Specific Perceived Knowledge (SPK), Perceived Intelligence (PI), Factuality (FCTY), and Willingness for Future Interaction (WFI). Statistically significant differences are marked with an asterisk (*). Error bars represent standard deviation.

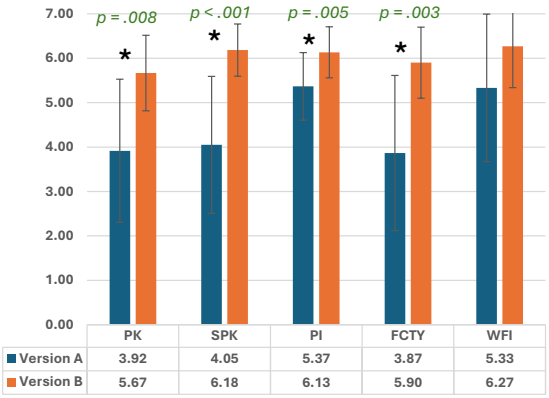


Fig. 4: Overall Partner Modeling Questionnaire (PMQ) [17] results comparing Version A and Version B: Communicative Competence and Dependability (CCD), Human-Likeness in Communication (HLC), and Communicative Flexibility (CF). Statistically significant differences are marked with an asterisk (*). Error bars represent standard deviation.

VII. Conclusions

This work presented an exploratory study on environment-aware ECAs empowered by VLMs, aiming at analyzing user's perception towards them across several dimensions. Thus, it outlined the integration of an EPM into conventional ECA architectures. The proposed pipeline comprised extracting image frames from the 3D environment via virtual cameras, processing these frames, and generating a SG representation of the VE using a VLM. The module was validated through a state-of-the-art-inspired use case, enabling the ECAs of a videogame application to exhibit environmental awareness.

The results obtained from within-subject experiments indicated a substantial enhancement of users' perception towards ECAs empowered with the devised EPM (Version B) in comparison to those without environmental knowledge (Version A) across various dimensions. Particularly, users perceived agents of Version B as more knowledgeable (PK, SPK), more intelligent (PI) and truthful in their responses (FCTY), as well as more competent in their communication skills (CCD). On the contrary, HLC, CF and WFI seemed to not vary between the two versions. Although preliminary, these findings provide a basis for further investigations into the importance of player perception towards ECAs for enhancing HAI. It should be noted that the study focused on perceived grounding rather than objectively measured grounding; no direct comparison between generated scene graphs and engine-level ground truth was conducted.

Despite the results, some limitations and challenges can be discussed. The proposed EPM currently supports only static VEs, as the SG is generated during initialization and remains unchanged throughout gameplay. This constrains the approach in dynamic contexts, where continuous SG recomputation would be necessary to maintain environmental fidelity and user immersion. Additionally, achieving both real-time performance and accurate SG generation remains an open issue. In particular, no comparison between generated scene graphs and engine-level ground truth was conducted in this study; therefore, the results specifically concern perceived grounding rather than objective scene understanding accuracy.

A further limitation concerns spatial orientation awareness. Although additional spatial predicates could, in principle, be incorporated into the SG, the present multi-view image-based method does not ensure reliable orientation information from the ECA's viewpoint. Hybrid approaches that combine VLM-based attribute extraction with engine-level techniques (e.g., ray casting or object tagging) could enhance spatial precision, but at the expense of scalability and increased integration complexity.

Future work will address these aspects by considering larger and more diverse samples, objective scene-graph validation against ground truth, dynamic environments with real-time updates, and hybrid approaches combining VLM perception with engine-level techniques for improved spatial precision.

Finally, the evaluation was conducted on a small sample and relied solely on users' subjective perceptions. While appropriate for an exploratory design, the limited sample size constrains the generalizability of the findings.

References

[1] Y. Tang, J. Situ, A. Y. Cui, M. Wu, and Y. Huang, "LLM integration in extended reality: A comprehensive review of current trends, challenges, and future perspectives," in *Proc. of the 2025 CHI Conf. on Human Factors in Computing Systems*, ser. CHI '25. ACM, 2025.

[2] F. R. Christiansen, L. N. Hollensberg, N. B. Jensen, K. Julsgaard, K. N. Jespersen, and I. Nikolov, "Exploring presence in interactions with LLM-driven NPCs: A comparative study of speech recognition and dialogue options," in *Proc. of the 30th ACM Symp. on Virtual Reality Software and Technology*, ser. VRST '24. ACM, 2024, pp. 1–11.

[3] F.-C. Yang, K. Duque, and C. Mousas, "The effects of depth of knowledge of a virtual agent," *IEEE Trans. on Visualization and Computer Graphics*, vol. 30, no. 11, pp. 7140–7151, 2024.

[4] Z. Li, H. Zhang, C. Peng, and R. Peiris, "Exploring large language model-driven agents for environment-aware spatial interactions and conversations in virtual reality role-play scenarios," in *Proc. of 2025 IEEE Conf. Virtual Reality and 3D User Interfaces (VR)*, 2025, pp. 1–11.

[5] H. Li, Z. Wang, W. Liang, and Y. Wang, "X's Day: Personality-driven virtual human behavior generation," *IEEE Trans. on Visualization and Computer Graphics*, vol. 31, no. 5, pp. 3514–3524, 2025.

[6] T. Zemskova and D. Yudin, "3DGraphLLM: Combining semantic graphs and large language models for 3D scene understanding," https://doi.org/10.48550/arXiv.2412.18450, 2025.

[7] M. Pan, A. Kitson, H. Wan, and M. Prpa, "ELLMA-T: an embodied llm-agent for supporting english language learning in social VR," in *Proc. of the 2025 ACM Designing Interactive Systems Conf.*, ser. DIS '25. ACM, 2025, p. 576594.

[8] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, and T. Gui, "The rise and potential of large language model based agents: A survey," https://doi.org/10.48550/arXiv.2309.07864, 2023.

[9] T. Rist, "Using a large language model to turn explorations of virtual 3D-worlds into interactive narrative experiences," in *Proc. of 2024 IEEE Conf. on Games (CoG)*, 2024, pp. 1–8.

[10] A. Hebri, H. R. Pavel, S. Nikanfar, F. Farahanipad, and F. Makedon, "Can virtual AI agents improve cognitive assessment? a virtual reality perspective," in *Proc. of the 18th ACM Int. Conf. on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '25. ACM, 2025, pp. 424–428.

[11] M. Konenkov, A. Lykov, D. Trinitatova, and D. Tsetserukou, "VR-GPT: Visual language model for intelligentvirtual reality applications," https://doi.org/10.48550/arXiv.2405.11537, 2024.

[12] R. Bayat, E. De Maio, J. Fiorenza, M. Migliorini, and F. Lamberti, "Exploring methodologies to create a unified VR user-experience in the field of virtual museum experiences," in *2024 IEEE Gaming, Entertainment, and Media Conf. (GEM)*, 2024, pp. 1–4.

[13] J. Zhu, R. Kumaran, C. Xu, and T. Hllerer, "Free-form conversation with human and symbolic avatars in mixed reality," in *Proc. of 2023 IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*, 2023, pp. 751–760.

[14] D. Calandra, F. G. Pratticò, and F. Lamberti, "Comparison of hands-free speech-based navigation techniques for virtual reality training," in *2022 IEEE 21st Mediterranean Electrotechnical Conf. (MELECON)*, 2022, pp. 85–90.

[15] M. A. Bravo, S. Mittal, S. Ging, and T. Brox, "Open-vocabulary attribute detection," in *Proc. of 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7041–7050.

[16] S. Häsler and P. Ackermann, "Spatial reasoner: A 3D inference pipeline for XR applications," https://doi.org/10.48550/arXiv.2504.18380, 2025.

[17] P. R. Doyle, I. Gessinger, J. Edwards, L. Clark, O. Dumbleton, D. Garaialde, D. Rough, A. Bleakley, H. P. Branigan, and B. R. Cowan, "The partner modelling questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners," *ACM Trans. Comput.-Hum. Interact.*, vol. 32, no. 4, pp. 1–33, 2025.