A Multisensory AI-based Framework for Accessible Navigation of Visually Impaired Users in Virtual Environments: Preliminary Results

Luiza M. F. Cintra[1] **, Elisa A. M. Oliveira** [1]**, Gustavo H. W. Barbosa** [1]**,**
Matheus D. Negrão [1]**, Valdemar V. G. Neto**[2]**,**
Rafael T. Sousa[1]**, Sofia L. C. Paiva** [2]**, Arlindo R. G. Filho**[1]

[1]Advanced Knowledge Center for Immersive Technologies
Goiânia – GO – Brazil

[2]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia, GO

Abstract. *Virtual reality has the potential to deliver highly immersive experiences, but for individuals with visual impairments, these environments often remain inaccessible and exclusionary. This paper introduces an AI-driven framework that redefines how such users interact with 3D virtual worlds. The system employs Vision-Language Models (VLMs) for real-time semantic scene understanding, translating visual information into auditory cues and haptic feedback. This multisensory approach allows users to perceive spatial layouts, recognize objects, and navigate with greater autonomy. By bridging the gap between visual content and non-visual perception, the framework turns virtual reality into a more inclusive, equitable, and engaging medium.*

1. Introduction

Virtual Reality (VR) has become a powerful tool for creating immersive environments that simulate real-world scenarios and foster experiential interaction [Creed et al. 2024]. Unlike traditional interfaces, VR promotes active engagement by allowing users to interact with dynamic, three-dimensional content in real time. Recent applications in diverse areas—such as training, language learning, entertainment, and STEM[1] exploration—demonstrate VR's potential to improve understanding, task performance, and motivation [Anjos et al. 2024]. However, most of these implementations rely heavily on visual interaction, limiting their accessibility and effectiveness for individuals with visual impairments [Creed et al. 2024].

The recent advancements in multimodal artificial intelligence (AI)—especially through Vision-Language Models (VLMs) that combine computer vision with natural language processing (NLP)—have opened new pathways for contextual understanding and interactive learning. These models are capable of interpreting visual scenes and generating or understanding text based on visual input, enabling more intuitive and responsive interactions in visually dynamic environments.This synergy between vision and language processing can play a key role in making digital content more accessible and intelligent [Mott et al. 2019].

[1]Science, Technology, Engineering and Mathematics

However, despite these technological advancements, significant accessibility barriers persist—especially for users with visual impairments. Common issues include: (1) the inability to receive timely or meaningful feedback; (2) difficulty in interpreting in-VR environment responses; and (3) challenges in providing input through conventional input devices [Yuan et al. 2011], [Heilemann et al. 2021]. These limitations highlight the need for inclusive design approaches that leverage multimodal AI not only to enhance interactivity, but also to ensure that users with diverse disabilities can fully engage with virtual environments.

The main contribution of this paper is a multisensory framework (a set of technological artifacts) that leverages generative AI—specifically Vision-Language Models (VLMs)—to support accessible navigation for visually impaired users within immersive virtual environments. The framework integrates VR technology with haptic and auditory modalities, translating visual stimuli into meaningful semantic information delivered through sound and touch. Unlike conventional VR systems that rely heavily on visual interaction, our approach is grounded in inclusive design principles, enabling non-visual exploration and interaction. As a proof of concept, we applied the framework in a structured virtual scenario (a classroom-like environment), demonstrating how it can support autonomous navigation and engagement with spatial and content-rich 3D environments in general.

The research objectives of this work are closely aligned with the proposed multisensory framework. Specifically, we aim to (i) investigate how Vision-Language Models (VLMs) can be effectively integrated into accessibility workflows to translate visual information into meaningful auditory and haptic cues; (ii) design and evaluate a proof-of-concept virtual classroom that allows visually impaired users to explore and interact autonomously using multimodal feedback; and (iii) derive inclusive design guidelines that can inform the development of future VR systems for non-visual navigation. By achieving these objectives, this study not only demonstrates the practical potential of the framework but also highlights its scalability toward real-world applications, such as wearable devices, ultimately promoting equitable participation in immersive environments for visually impaired individuals.

This paper is structured as follows: Section 2 contextualize the tools adopted to create the framework and related work; Section 3 outlines the research method; Section 4 explores the development of the immersive environment with auditory and haptic feedback and Section 5 discusses the final remarks.

2. Related Works

An exploratory literature review reveals several studies approaching accessibility in VR environments. One notable example is Empath-D [Kim et al. 2018], a system developed to help application designers evaluate the usability of their mobile apps from the perspective of users with disabilities. Our approach, in contrast, shifts the perspective to the actual user experience of individuals with visual impairments by using VLMs to interpret scenes and guide navigation in VR.

Another related work is SeeingVR [Zhao et al. 2019]. It proposes a set of 14 tools to improve the accessibility of virtual reality applications for users with low vision, through visual and auditory augmentations such as magnification, contrast enhancement,

edge highlighting, and text-to-speech. In contrast, our framework is aimed at visually impaired users, who cannot benefit from visual augmentations. Rather than enhancing visual content, our approach offers non-visual multisensory guidance—allowing users to explore and navigate immersive environments through auditory cues and haptic feedback derived from semantic scene understanding.

A distinct line of research explores how users with motor limitations interact with virtual products [Grande et al. 2025]. That work focuses on adapting interactions to assist in object manipulation. On the other hand, our framework addresses a different accessibility gap: non-visual navigation within immersive environments. Our contribution lies in enabling visually impaired users to move through complex 3D spaces autonomously, using a combination of scene understanding through VLMs and multimodal (audio-haptic) feedback.

While existing works address accessibility in VR environments across various disabilities, to the best of our knowledge, none combine the use of multisensory interaction, semantic scene interpretation, and generative AI to support visually impaired users in immersive environment navigation, as proposed in our work.

3. Research Method

For the settlement of the framework, we established a method structured into well-defined steps. Firstly, we conducted an ***Ad-hoc* literature review (Step 1)** searching for studies on (i) virtual environments with accessibility for visual disorders, (ii) strategies to support non-visual navigation and exploration in immersive 3D spaces and (iii) integrating AI generative for accessibility in virtual environments. Based on the literature review, several multimodal models commonly used in computer vision contexts were identified. Among them, GPT-4o was found to be the most suitable for the purposes of this application [Wen et al. 2024], as it is one of the most widely used multimodal models for tasks involving image description and visual question answering (VQA) [Li et al. 2025]. Following the selection of the model, a **series of input tests were conducted by authors to evaluate the quality of its responses (Step 2)**. Examples of these prompts are described below.

- "*What do you observe in the image?*",
- "*Take me to the nearest chair so I can attend the class*.",
- "*Take me to the bookshelves.*"

After validating the model's responses using specific questions based on images from the virtual environment, as shown in Figure 1, developed for this application (as well as assessing its ability to guide users toward different objects within those images), the study advanced to the **development of the complete framework in Unreal Engine (Step 3)**, integrated with the selected vision-language model. The preliminary result of this research is the definition of the multisensory framework, represented in Figure 2, rather than the implementation of the virtual scenario itself. The framework is organized into two complementary processing flows. In the auditory flow, the user issues a spoken query about the virtual environment (e.g., "What can I perceive around?"). The speech is transcribed by Whisper, while a screenshot of the VR environment is simultaneously captured, encoded, and sent to the Vision-Language Model (GPT), which interprets the scene and generates a textual response. This response is then converted into audio and

Figure 1. Scene of the developed scholar environment.

delivered back to the user through Text-to-Speech synthesis. In parallel, the haptic flow provides non-visual feedback based on spatial interactions within the environment: the user's geolocation triggers collision detection in Unreal Engine, and when an object is approached or touched, the controller delivers haptic signals that communicate proximity or contact. Together, these two flows form the framework's core structure, enabling visually impaired users to navigate and interact with immersive environments through semantic audio descriptions and tactile cues.

The future stages **(Step 4 and Step 5)** will focus on the evaluation and refinement of the framework. Step 4 will involve user testing with visually impaired participants to assess usability, accessibility, and effectiveness of the multisensory feedback. Step 5 will integrate the system with wearable devices (e.g., smart glasses equipped with cameras) to explore real-world applications beyond VR. These stages will be essential to validate the framework's practical impact and consolidate its contributions to inclusive design.

4. Description of the Developed Prototype

The assessment of the proposed framework was conducted through its instantiation in a three-dimensional virtual environment implemented in Unreal Engine 5. For experimental control, the environment was configured with a classroom layout, selected exclusively as a structured scenario that provides a variety of spatial references and obstacles for evaluating navigation and orientation mechanisms. It is important to note that this configuration serves only as a test environment: the framework is independent of domain semantics and can be deployed in any virtual 3D environment. Within this configuration, the environment reproduces a realistic classroom, including desks, a whiteboard, bookshelves, a backpack on the floor, ambient classroom sounds, and spatialized audio cues. To enable natural interaction, the system integrates GPT-4o, a lightweight vision-language model capable of interpreting the user's current visual context through high-resolution rendered frames (1541×900 pixels) and, in this case, generating the corresponding output response.

When a user issues a voice query, the speech is processed using a Speech-to-Text (STT) module, called Whisper, which transcribes the input (converting a WAV audio file to text) and forwards it to GPT-4o for contextual understanding. The model generates a semantic response based on the visual content of the current scene, which is then synthesized into audio using a Text-to-Speech (TTS) engine. Additionally, haptic feedback is provided through a wearable device to reinforce spatial orientation and object recognition within the scene. This multimodal system enables non-visual navigation and learning, offering an inclusive educational experience that combines auditory, tactile, and

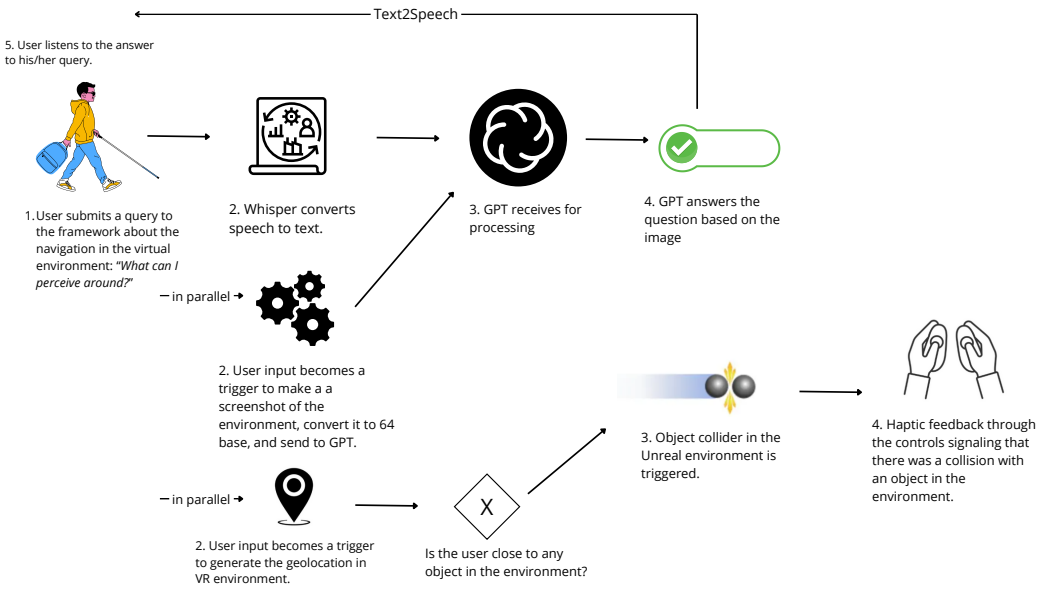

Figure 2. System architecture for AI-based interactive assistance integrated with Unreal Engine. The diagram illustrates the workflow of the proposed system. User input is processed by Whisper-Tiny for transcription and combined with a scene screenshot and geolocation data from Unreal. These inputs are sent to GPT-4o, which interprets the user's intent. The output is used in two ways: generating speech feedback via a text-to-speech module, and triggering object collisions in the virtual environment to provide force feedback and additional audio cues.

semantic information in real-time. In the implementation, force feedback is delivered via Unreal Engine's haptic system, which allows developers to convey physical forces through programmable vibration patterns on supported devices. This system utilizes Force Feedback Effect assets composed of intensity curves across multiple motor channels, enabling precise control over the strength and duration of vibrations. These effects can be triggered dynamically at runtime, and can be spatialized to reflect the user's proximity to objects or points of interest. By aligning vibration intensity with navigational cues, the system provides an additional non-visual layer of spatial information that enhances user awareness and autonomy in immersive environments.

Upon initialization, the application provides an audio-based spatial description of the environment, including the location and characteristics of relevant objects. Based on this information, users are able to navigate the space and interact with predefined interactive elements. For instance, when a user requests to move toward the bookshelf, the system generates a directional response that guides the user through the virtual space. Navigation is performed via VR controllers, and the user may request updated guidance at any moment. Haptic feedback is integrated into the system, with vibration intensity increasing as the user approaches the target location. Upon reaching the specified object (e.g., the bookshelf), maximum vibration indicates arrival and enables interaction—such as collecting a book to complete an exploratory task. From any point within the environment, users can request further guidance, relying on a combination of auditory and haptic cues to support independent orientation and mobility.

5. Final Remarks

This work presented the development of a multimodal framework designed to support visually impaired users in navigating three-dimensional virtual environments. By integrating high-fidelity VR technologies with vision-language models, auditory narration, and haptic feedback mechanisms, the framework enables users to explore and interact with virtual spaces without relying on visual input. The system leverages GPT-4o for contextual understanding of rendered scenes and generates adaptive responses based on user queries, allowing real-time guidance through spatialized descriptions and vibration-based cues. The use of Unreal Engine's programmable force feedback system further enhances spatial orientation, reinforcing proximity awareness and object recognition. Although the experimental evaluation was conducted in a classroom-shaped environment, the framework itself is domain-independent and can be applied to any virtual 3D scenario that requires accessible navigation.

Despite the promising results, this work is still in its early stages and offers several opportunities for future development. Planned next steps include a comprehensive experimental validation of the framework with users presenting different levels of visual impairment. This evaluation will combine quantitative metrics—such as navigation efficiency, accuracy, task completion time, frequency of orientation requests, and navigation errors—with qualitative data collected through interviews, user observations, and standardized instruments such as the System Usability Scale (SUS). The results of these studies will guide refinements in interaction strategies, including the dynamic adjustment of feedback intensity and the adaptation of guidance behavior according to user profiles. In parallel, future iterations will extend the application of the framework to a broader range of 3D scenarios beyond the current controlled testbed, with the goal of confirming its generality and robustness across diverse virtual environments.

Acknowledgments

This work has been fully/partially funded by the project Research and Development of Algorithms for Construction of Digital Human Technological Components supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

References

Anjos, F. E. V. d., Martins, A. d. O., Rodrigues, G. S., Sellitto, M. A., and Silva, D. O. d. (2024). Boosting engineering education with virtual reality: An experiment to enhance student knowledge retention. *Applied System Innovation*, 7(3):50.

Creed, C., Al-Kalbani, M., Theil, A., Sarcar, S., and Williams, I. (2024). Inclusive ar/vr: accessibility barriers for immersive technologies. *Universal Access in the Information Society*, 23(1):59–73.

Grande, R., Albusac, J., Herrera, V., Monekosso, D., De Los Reyes, A., Vallejo, D., and Castro-Schez, J. (2025). Enhancing hand interactions and accessibility in virtual reality environments for users with motor disabilities: A practical case study on vr-shopping. *IEEE Access*.

Heilemann, F., Zimmermann, G., and Münster, P. (2021). Accessibility guidelines for vr games-a comparison and synthesis of a comprehensive set. *Frontiers in Virtual Reality*, 2:697504.

Kim, W., Choo, K. T. W., Lee, Y., Misra, A., and Balan, R. K. (2018). Empath-d: Vr-based empathetic app design for accessibility. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 123–135.

Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., and Shi, G. (2025). A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606.

Mott, M., Cutrell, E., Franco, M. G., Holz, C., Ofek, E., Stoakley, R., and Morris, M. R. (2019). Accessible by design: An opportunity for virtual reality. In *2019 IEEE international symposium on mixed and augmented reality adjunct (ISMAR-adjunct)*, pages 451–454. IEEE.

Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., Ma, T., Li, Y., Xu, L., Shang, D., et al. (2024). On the road with gpt-4v (ision): Explorations of utilizing visual-language model as autonomous driving agent. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Yuan, B., Folmer, E., and Harris Jr, F. C. (2011). Game accessibility: a survey. *Universal Access in the information Society*, 10(1):81–100.

Zhao, Y., Cutrell, E., Holz, C., Morris, M. R., Ofek, E., and Wilson, A. D. (2019). Seeingvr: A set of tools to make virtual reality more accessible to people with low vision. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.