Journal on Interactive Systems, 2026, 17:1
ISSN: 2763-7719 • doi: 10.5753/jis.2026.7402

Research Paper

A Semantic Multimodal Architecture for Accessible Navigation in Immersive Virtual Environments

Luiza Cintra [Federal University of Goiás | *cintraluiza@discente.ufg.br*]
Elisa Ayumi [Federal University of Goiás | *ayumi@discente.ufg.br*]
Gustavo Webster [Federal University of Goiás | *guga.webster@gmail.com*]
Andressa Bastos [Federal University of Goiás | *andressa.bastos08@gmail.com*]
Valdemar V. Graciano-Neto [Federal University of Goiás | *valdemarneto@ufg.br*]
Sofia Costa Paiva [Federal University of Goiás | *sofialarissa@ufg.br*]
Rafael Sousa [Federal University of Goiás | *rafael.sousa@ufmt.br*]
Arlindo Galvão [Federal University of Goiás | *arlindogalvao@ufg.br*]

Institute of Informatics, Universidade Federal de Goiás, Alameda Palmeiras, s/n - Chácaras Califórnia, Goiânia, GO, 74690-900, Brazil.

Abstract. Accessible interaction in immersive virtual environments remains largely constrained by visually driven design paradigms, limiting meaningful engagement for users with visual impairments. Existing approaches to VR accessibility often rely on application-specific adaptations, low-level sensory cues, or vision-centered augmentations, resulting in fragmented and non-scalable solutions. This paper introduces a conceptual architecture for a semantic multimodal interface that mediates access to immersive virtual environments through non-visual representations. The architecture operates as a reusable interaction layer that translates visual information into coherent auditory and haptic feedback grounded in semantic understanding, supporting users across the full spectrum of visual impairment, from total blindness to low vision. Central to the architecture is the use of Vision Large Language Models as a semantic mediation layer, enabling dynamic interpretation of visual scenes and context-aware multimodal feedback independent of application logic. The interface is designed as a modular architecture that can be integrated into diverse VR applications, decoupling accessibility mechanisms from scenario-specific implementations. The architecture is instantiated in heterogeneous immersive environments to illustrate its generality and architectural feasibility. By treating accessibility as a foundational interface concern rather than an add-on feature, this work contributes a design-oriented perspective for scalable, inclusive interaction in immersive virtual environments.

Keywords: Virtual Reality, Accessibility, Visual Impairment, Haptic Feedback, Vision-Language Models

Edited by: Cleber Gimenez Corrêa | **Received:** 15 January 2026 • **Accepted:** 28 August 2026 • **Published:** 08 September 2026

1 Introduction

Virtual Reality (VR) has increasingly established itself as a robust technological medium for constructing immersive digital environments capable of simulating real-world situations and supporting experiential forms of interaction Delaney and Biocca [1995]. Differently from conventional two-dimensional interfaces, VR environments enable users to engage directly with spatially situated, three-dimensional content that evolves dynamically in response to user actions. This characteristic fosters embodied interaction, where perception, movement, and cognition are closely coupled within the virtual space Rubio-Tamayo *et al*. [2017]. As a result, VR has been widely adopted across multiple domains, including professional training, second-language acquisition, entertainment, and STEM[1] education, with empirical studies reporting improvements in conceptual understanding, task efficiency, engagement, and intrinsic motivation Anjos *et al*. [2024]. Despite these advances, the majority of VR applications remain predominantly vision-centered, relying on graphical cues and visual feedback as the primary means of interaction. This design paradigm substantially constrains accessibility and diminishes the effectiveness of VR experiences for users with visual impairments, who may be unable to perceive or interpret essential visual information Creed *et al*. [2024].

Visual impairment represents a heterogeneous set of sensory conditions, spanning from complete blindness to various forms of low vision (LV). Individuals with total blindness typically rely on non-visual sensory channels—most notably auditory and tactile perception—to navigate and understand their environment. In contrast, people with low vision preserve residual visual capacity, which may manifest as blurred or distorted vision, tunnel vision, reduced visual acuity, or diminished contrast sensitivity, often resulting from conditions such as cataracts or glaucoma (Figures 1 and 2 illustrate simulated visual perception under these conditions). These impairments impose significant challenges for spatial orientation, object recognition, and safe navigation, frequently limiting personal autonomy, independent mobility, and participation in daily activities Leat *et al*. [1999]. Importantly, while assistive tools such as white canes and guide dogs are well-established solutions for individuals with total blindness, they do not fully address the needs of LV users, who must continuously integrate partial visual input with auditory and tactile cues. This hybrid perceptual strategy requires adaptive assistive systems capable of dynamically balancing multimodal information rather than replacing vision entirely Pur *et al*. [2023].

Recent progress in multimodal artificial intelligence (AI) has created new opportunities to address these accessibility

[1] Science, Technology, Engineering and Mathematics

Figure 1. Simulation of vision with cataract: the scene appears globally blurred, with reduced contrast and desaturated colors, creating a hazy, fog-like effect that obscures fine details.

Figure 2. Simulation of vision with glaucoma: progressive loss of peripheral vision results in a "tunnel vision" effect, where only the central region remains relatively clear while the periphery is darkened or absent.

challenges. In particular, Vision Large Language Models (VLLMs)—which integrate computer vision techniques with natural language processing (NLP)—enable systems to analyze visual scenes, extract semantic meaning, and generate descriptive or instructional language grounded in visual input. These models support bidirectional interaction, allowing users not only to receive contextualized explanations of visual content but also to issue natural-language queries about their surroundings Shen *et al*. [2022]. By coupling perceptual interpretation with linguistic reasoning, VLLMs facilitate more adaptive, context-aware, and intelligible interactions in visually rich environments, representing a significant step toward intelligent accessibility solutions Mott *et al*. [2019].

Nevertheless, despite the maturation of VR hardware and the emergence of multimodal AI models, substantial accessibility barriers persist for users with visual impairments. Prior research has identified recurring issues, including delayed or insufficient feedback during interaction, difficulties in interpreting environmental responses within immersive spaces, and limited usability of standard input devices such as handheld controllers or gesture-based interfaces Heilemann *et al*. [2021]; Yuan *et al*. [2011]. These obstacles often result in increased disorientation and frustration, ultimately discouraging sustained use of VR systems by visually impaired users. Such findings underscore the necessity of inclusive design approaches that move beyond visual substitution and instead leverage multimodal AI to support meaningful, equitable interaction across diverse sensory capabilities.

In response to these challenges, the main contributions of this paper are:

- **A multisensory accessibility architecture for immersive VR based on VLLMs.** We propose a domain-agnostic interaction architecture that integrates VLLMs with auditory and haptic feedback to translate visual scene information into semantically meaningful, non-visual representations, enabling accessible navigation and interaction for users with visual impairments.
- **A reusable interaction layer decoupled from application-specific logic.** The proposed architecture abstracts visual information into multimodal representations that can be reused across different VR applications, allowing accessibility features to be integrated independently of the underlying environment or task domain.
- **The design and implementation of two VR scenarios.** We developed and evaluated the architecture in two contrasting immersive environments—a classroom and a hospital reception hall—to examine its applicability across different spatial configurations and interaction demands.
- **An exploratory user study with visually impaired participants.** We conducted a study with blind and low-vision users performing navigation and interaction tasks using only auditory and haptic feedback. The results provide initial empirical evidence supporting the feasibility of the proposed approach.
- **An analysis of multimodal interaction in VLLM-driven VR systems.** We discuss how the combination of generative AI and multisensory feedback impacts spatial understanding, navigation behavior, and user experience, as well as current limitations.

Finally, the remainder of this paper is structured as follows. Section 2 reviews the background and related work. Section 3 presents the research method adopted to develop the pipeline. Section 4 discusses the proposed multisensory accessibility architecture, detailing its design principles. Section 5 describes the interface architecture and the implementation of the architecture within immersive VR environments, including the development of the classroom and hospital reception scenarios. Section 6 outlines the evaluation and user study protocol. Section 7 discusses the findings, limitations, and broader implications, including ethical and societal considerations. Finally, Section 8 concludes the paper and highlights directions for future work.

1.1 Ethical issues

The project was registered and approved by the Institutional Ethics Committee (Protocol 88442725.6.0000.5083; 25.07.2025). The research involved user testing sessions to evaluate the proposed accessibility architecture for immersive virtual environments. All procedures followed established ethical guidelines to ensure informed consent, participant privacy, voluntary participation, and the minimization of potential physical or psychological risks.

Given the involvement of participants with visual impairments, particular attention was devoted to accessibility and safety during the experimental sessions, including clear task instructions, assisted setup when needed, and continuous monitoring to prevent discomfort or disorientation in immersive environments. Participants were informed about the experimental nature of the system and its reliance on AI-generated outputs, including the possibility of inaccuracies in system responses.

The architecture relies on GPT-4o, a commercially operated model accessed via API, which implies that rendered scene data—including screenshots of the virtual environment—are transmitted to external servers for processing. The use of VLLM is framed as assistive and mediating, with explicit acknowledgment of potential biases and limitations in generated content, and its use is disclosed in accordance with the Brazilian Computer Society's Code of Conduct and COPE's core practices. Future iterations of the architecture should explore locally deployable open-source VLLMs as a privacy-preserving alternative for sensitive institutional contexts such as healthcare and education. Also, personal data collected during the study—including demographic information and questionnaire responses—were handled in strict accordance with the Lei Geral de Proteção de Dados (LGPD), ensuring that all data were anonymized prior to analysis, stored securely, and used exclusively for the purposes outlined in the informed consent form. No personal data were shared with third parties or retained beyond the scope of this research.

A further ethical consideration concerns the potential impact of model errors on users who rely exclusively on non-visual feedback. Unlike sighted users who can independently verify environmental information, visually impaired users may be more vulnerable to the consequences of hallucinations or spatially misleading outputs. This asymmetry reinforces the importance of transparency: users must be clearly informed of the AI-generated nature of descriptions and the possibility of inaccuracies, and the interface should provide mechanisms for requesting repeated or alternative guidance. Additionally, the architecture should be designed to scaffold rather than substitute user agency, supporting the gradual development of independent spatial reasoning rather than fostering over-reliance on AI-mediated perception.

2 Background and Related Work

Effectively navigating and interacting within immersive digital environments presents a significant accessibility challenge for individuals across the entire spectrum of visual impairment Real and Araujo [2019]. While individuals with total blindness depend exclusively on non-visual sensory channels, such as haptics and audition, to build a mental model of their surroundings, the needs of users with LV are uniquely complex. Users with LV strive to leverage their residual vision, but this ability is often compromised by conditions like blur, tunnel vision, or poor contrast sensitivity. In visually dense or dynamic scenes, relying solely on impaired vision can lead to cognitive overload, navigational errors, and a diminished sense of presence Fryer [2013]. Therefore, there is a critical need for robust non-visual feedback that can supplement and disambiguate the limited visual information LV users perceive, just as it must substitute vision entirely for those who are totally blind.

VR environments intensify this challenge. By design, they are powerful immersive platforms, but their efficacy is overwhelmingly tied to high-fidelity visual feedback. This paradigm inherently excludes users with total blindness and can prove equally frustrating for LV users, who may find their residual vision insufficient to reliably interpret the complex spatial layouts and interactive elements common in VR. This highlights a critical gap in assistive technology design: the need for a unified architecture that can deliver meaningful, contextual information through non-visual means Oliveira *et al*. [2023]; Ortiz-Escobar *et al*. [2023], thereby supporting both the complete substitution of vision for blind users and the powerful supplementation of vision for LV users. These prior works validate the foundational use of non-visual feedback for VR navigation but also reveal a reliance on pre-scripted cues or low-level sensory data. In the current study, we concentrate on how generative AI—specifically VLLMs—can bridge this gap. We investigate how translating the visual scene into rich, semantic information supports not only mobility and obstacle avoidance but also meaningful object interaction and autonomous exploration for users across the visual impairment spectrum.

A significant contribution aimed at LV users is SeeingVR Zhao *et al*. [2019], a toolkit comprising 14 augmentations for existing VR applications. The system offers a suite of tools that users can combine and customize, including visual enhancements like magnification, edge enhancement, and brightness adjustments, as well as supplementary audio feedback. The architecture is designed for flexibility, allowing developers to integrate the tools via a Unity toolkit or enabling users to apply a subset of them post-hoc to an application without developer intervention. While SeeingVR provides a robust and effective solution for its target audience, its design is fundamentally centered on augmenting and enhancing a user's residual vision. As such, its tools do not extend support to individuals with total blindness. Furthermore, its methodology focuses on direct sensory augmentation rather than on semantic interpretation of the environment. In contrast, our current study leverages VLLMs to translate complex visual scenes into rich, contextual information delivered through non-visual modalities. This approach allows our architecture to address the needs of both totally blind and LV users, focusing on providing a deeper, semantic understanding of the virtual space rather than solely enhancing visual perception.

Beyond specific applications, broader literature reviews have begun to map the emerging landscape of AI-driven accessibility in VR. For instance, Liu *et al*. [2024] highlight the risk of a widening accessibility gap as VR becomes more integrated into critical areas like education and employment,

potentially disadvantaging users with disabilities. Their work surveys the integration of AI into VR systems as a promising pathway to mitigate this risk, particularly for blind and low vision users. The authors identify a significant gap in the existing literature: a lack of comprehensive reviews that synthesize the benefits, challenges, and future opportunities of using AI to make VR accessible. By providing a structured overview of current research, the review serves to consolidate the field and underscores the need for more focused contributions. This call for novel research directly validates the direction of our study. Our work responds to the opportunities identified in such reviews by proposing and evaluating a concrete architecture that operationalizes the use of advanced AI to address the persistent accessibility challenges within immersive environments.

Another approach focused on visual augmentation is VRiAssist Masnadi *et al*. [2020], a tool that utilizes eye-tracking to assist users with visual impairments in VR. The core innovation of VRiAssist is its ability to project visual corrections dynamically onto the specific area of the user's gaze, directly addressing the affected region of their retina. The system includes several tools, such as distortion correction to counteract retinal irregularities, color and brightness adjustments to improve contrast, and an adjustable magnifier. The VRiAssist system represents a sophisticated method for real-time visual enhancement, leveraging eye-tracking to deliver highly targeted corrections. However, similar to other augmentation-focused tools, its design is exclusively centered on users who retain a degree of functional vision (low vision) and does not apply to individuals with total blindness. Our research diverges from this approach by not attempting to correct or enhance the user's visual perception directly. Instead, our architecture focuses on translating the entire visual scene into a cohesive, non-visual experience through semantic audio and haptic feedback powered by AI, making it a viable solution for a broader spectrum of visual impairments.

Finally, one more relevant approach to non-visual interaction in VR is presented in Zhao *et al*. [2018], a haptic cane controller that replicates white cane interactions in VR through physical resistance, vibrotactile feedback, and spatialized audio, enabling users to transfer existing cane navigation skills into virtual spaces. While Canetroller represents a compelling proof of concept for skill-based navigation, its design is grounded in the physical metaphor of the white cane, which presupposes familiarity with cane use and addresses primarily the needs of users with total blindness. In contrast, the architecture proposed in this paper does not rely on prior assistive tool experience or a specific interaction metaphor. Instead, it employs VLLMs to generate semantic, context-aware descriptions of the virtual environment on demand, complemented by proximity-based haptic cues that are independent of any physical analogue. This approach supports both individuals with total blindness and those with LV, accommodating a broader spectrum of perceptual strategies without requiring users to map real-world motor skills onto virtual interaction.

The challenges of interpreting complex 3D environments are particularly pronounced for users with permanent visual impairments. These users can experience cognitive overload when navigating data-rich or unfamiliar virtual environments Freire *et al*. [2023], highlighting the need for systems that offload visual information to other sensory channels. Providing auditory summaries of a scene or haptic cues for navigation can significantly support spatial understanding and task performance. A central theme emerging from the literature is the importance of developing systems that translate complex visual data into meaningful, actionable information, making virtual environments more accessible and navigable. Prior work Giudice [2018]; Kane *et al*. [2011] indicates a clear trajectory away from simple, low-level alerts (e.g., proximity vibrations, directional beeps) and towards richer, contextual representations of the environment.

Our study builds upon these prior works by proposing a architecture that directly addresses this need for semantic interpretation. This approach allows us to investigate how a multisensory experience can also facilitate autonomous exploration and meaningful interaction for users with both total blindness and low vision Fernandes *et al*. [2024], empowering them with a deeper understanding of the virtual space.

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 models for accessibility in virtual environments. Regarding topic (i), the review identified a consistent pattern in the literature: immersive VR systems remain predominantly designed around visual interaction paradigms, which systematically exclude users with total blindness and impose significant challenges on those with LV Creed *et al*. [2024]; Heilemann *et al*. [2021]. Studies in this area documented recurring difficulties in spatial orientation, object recognition, and independent navigation, as well as the distinct needs of LV users, whose reliance on residual vision requires supplementary rather than substitutive feedback strategies Zhao *et al*. [2019]; Real and Araujo [2019]. Research across diverse socio-economic contexts has further highlighted accessibility as a matter of social equity, examining how technological advancements in VR can inadvertently become barriers for underserved populations Silveira *et al*. [2024]; Fernandes *et al*. [2024]. For topic (ii), reviewed works examined non-visual interaction techniques employed in immersive 3D environments, including spatialized audio, proximity-based haptic cues, and multimodal combinations Giudice [2018]; Yuan *et al*. [2011]. A recurring limitation across these approaches was their reliance on pre-scripted, low-level sensory mappings that constrain adaptability and fail to convey higher-order semantic information about the environment Ortiz-Escobar *et al*. [2023]; Fryer [2013]. Guidelines proposed for the design of accessible interfaces for users with visual impairments Silva *et al*. [2019] further reinforced the need for adaptive, context-aware solutions rather than fixed sensory substitution mechanisms. For topic (iii), the review surveyed studies on the integration of generative AI models for accessibility in virtual environments, with particular focus on identifying which models were most applicable to the demands of real-time scene interpretation and natural language generation in immersive contexts Shen

et al. [2022]; Mott *et al*. [2019]. Several multimodal models commonly used in computer vision tasks were identified and compared, including their capacity for image description, visual question answering, and context-sensitive language generation Liu *et al*. [2023]. Among them, GPT-4o emerged as the most suitable for the purposes of this application, given its established performance in tasks involving visual scene understanding and its ability to generate semantically coherent, adaptive responses grounded in visual input Kuang *et al*. [2025]; Wen *et al*. [2024]; Li *et al*. [2025b]. Following the selection of the model, a **series of input tests were conducted by the authors to evaluate the quality of its responses (Step 2)**. These tests were conducted by submitting static screenshots captured from both virtual environments directly to GPT-4o, independently of the Unreal Engine integration. The user inputs were designed to cover two complementary dimensions of the model's capability: its ability to produce semantically coherent descriptions of the visual scene, and its capacity to generate actionable navigational guidance toward specific objects. For each input, the authors evaluated whether the model's response accurately reflected the content of the image, whether the navigational instructions were spatially consistent with the depicted environment, and whether any hallucinations or object misidentifications occurred. The model was initialized with a system prompt defining its role as a non-visual navigation assistant, as shown in Figure 3. Figure 4 presents a concrete example of this validation process, illustrating a user query submitted alongside a virtual environment screenshot and the corresponding response generated by the model.

Additional examples of user inputs employed during this validation phase are presented below, spanning queries about general scene description, object localization, and goal-oriented navigational guidance:

```
Scene description:
  "What do you observe in the image?"
  "Describe the room I am currently in."
  "What objects are around me?"

Spatial orientation:
  "What is in front of me?"
  "What is to my left?"

Goal-oriented navigation:
  "Take me to the nearest
  chair so I can attend the class."
  "Take me to the bookshelves."
  "Guide me to the bathroom."
  "Where is the exit?"

Object interaction:
  "Is there a chair I can sit on nearby?"
  "Can I interact with anything in
  front of me?"
  "What is on the desk?"
```

After validating the model's responses through targeted queries, the research advanced to the **development of the complete architecture in Unreal Engine (Step 3)**, integrated with the selected VLLM. At this stage, the primary outcome of

```
ASSISTIVE GUIDE PROMPT

You are a highly trained assistive guide helping visually impaired individuals
navigate safely through indoor environments such as rooms, corridors, and offices.

Your task is to interpret the image and respond based on the user's question
or request provided in the transcription below.

IMPORTANT:
1. If the user says something like "describe the environment" or
   "what am I seeing," do not provide navigation instructions.
   Only describe the environment in as much detail as possible.
2. NEVER ask the user for more information. Work only with what
   has been provided and infer the best possible guidance.
3. Always consider the position and arrangement of objects and
   provide clear, tactile-oriented guidance when necessary.

FORMAT FOR GUIDANCE:
- Use a maximum of 4 short, well-detailed sentences.
- Act as a guide: the user needs spatial orientation to move safely.
- Use expressions such as: "follow along the right wall",
  "when you feel the table with your left hand, turn right".
- Always assume the user is positioned at the center of the image.
- Be fully consistent with what is shown in the image.
```

Figure 3. System prompt used to configure GPT-4o as a non-visual navigation assistant, defining its role, behavioral constraints, and response format prior to user interaction.

the research is not the validation of a specific virtual scenario, but the definition and implementation of a reusable multisensory architecture capable of mediating access to immersive environments for users with visual impairments. The architecture is organized around two complementary processing flows, which together constitute its core interaction architecture. In the auditory flow, the user issues a spoken query about the virtual environment (e.g., "What can I perceive around me?"). This input is transcribed using a Speech-to-Text module (Whisper-Tiny), while a snapshot of the current VR scene is simultaneously captured, encoded, and forwarded to the VLLM (GPT-4o). The model interprets the visual scene and generates a structured semantic response, which is subsequently converted into audio via Text-to-Speech synthesis and delivered to the user. This flow enables the transformation of visual content into meaningful auditory descriptions grounded in the current spatial context. In parallel, the haptic flow provides continuous non-visual feedback based on spatial interaction within the virtual environment. The user's position and movement are monitored through Unreal Engine's collision and proximity detection mechanisms. As the user approaches or interacts with virtual objects, programmable force-feedback effects are triggered on the controller, conveying proximity, direction, or contact through tactile cues. This haptic channel complements auditory feedback by supporting spatial orientation and interaction without requiring explicit verbal queries.

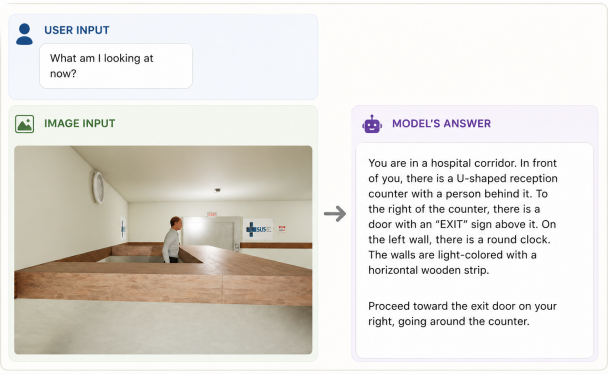

Figure 4. Representative validation example: a user query submitted alongside a screenshot of the virtual environment and the corresponding response generated by GPT-4o.

Together, these two flows form the conceptual and architectural foundation of the proposed architecture, enabling visually impaired users to navigate and interact with immersive environments through semantically mediated audio descriptions and tactile signals rather than direct visual perception, as demonstrated in Figure 5.

To demonstrate the feasibility and generality of the architecture **(Step 4), it was instantiated within two distinct immersive VR scenarios**: a classroom and a hospital reception hall. These scenarios were selected as illustrative cases representing environments with different spatial structures and interaction characteristics, rather than as experimental testbeds. Their purpose is to exemplify how the same architecture can be embedded within heterogeneous Unreal Engine applications, supporting accessibility across diverse immersive contexts without application-specific customization.

Building upon the architectural foundations established in the previous steps, the research culminates in a **conceptual analysis and synthesis of design implications (Step 5)**, focusing on how the proposed architecture addresses identified accessibility gaps, supports inclusive interaction, and enables extensible integration within immersive systems. Rather than reporting empirical performance metrics, this phase consolidates architectural insights and design considerations that inform the development of accessible, multimodal interfaces for virtual reality. The complete research process is summarized in the methodological flowchart presented in Figure 6.

To empirically validate the proposed architecture, we conducted a **user study with participants with visual impairments (Step 6)**. The study involved individuals with total blindness and LV across both immersive scenarios, who were asked to complete goal-oriented navigation tasks guided exclusively by the multimodal feedback generated by the system. Both quantitative data—such as task completion times and number of collisions—and self-reported responses were collected through structured questionnaires to evaluate the architecture's impact on spatial awareness, user confidence, and sense of presence.

Finally, **the experimental results were analyzed and synthesized (Step 7)**, combining quantitative metrics and subjective evaluations to derive empirical evidence of the architecture's feasibility and to identify opportunities for refinement. This phase consolidates both the technical and experiential dimensions of the evaluation, informing future iterations of the interface design. The complete research process is summarized in the methodological flowchart presented in Figure 6.

4 Design Rationale and Conceptual Architecture

The design of accessible immersive interfaces for users with visual impairments requires more than incremental adaptations of visually driven systems Bendaly Hlaoui *et al*. [2019]. As discussed in the previous section, current VR accessibility approaches remain fragmented, often limited to application-specific solutions, low-level sensory cues, or vision-centered augmentations. These limitations point to a broader design challenge: enabling meaningful interaction and spatial understanding in immersive environments without assuming visual primacy.

One way to address this challenge is to reframe accessibility as a problem of semantic mediation rather than sensory substitution. Creed *et al*. [2024] systematically identified and mapped interaction barriers across a spectrum of impairments in AR/VR environments—including visual, physical, cognitive, and auditory disabilities—highlighting the pervasive nature of accessibility gaps in immersive technologies. Building upon this characterization of the problem space, the present architecture proposes a concrete architectural response by conceiving accessibility as a reusable interaction layer that can be integrated across different VR applications and domains. Its goal is to support users across the full spectrum of visual impairment—from total blindness to LV—by translating visual information into coherent auditory and haptic feedback grounded in semantic understanding, thereby promoting equitable access to immersive technologies regardless of sensory ability Dudley *et al*. [2023]; Rodrigues *et al*. [2025].

This architecture is informed by principles of inclusive and universal design Story [2001]; Persson *et al*. [2015] and by prior research on multisensory interaction and cognitive accessibility Giudice [2018]; Fernandes *et al*. [2019]. Recognizing that visual impairment is not a binary condition, the design emphasizes adaptability and user autonomy, enabling users to construct mental models of virtual environments through complementary sensory channels rather than relying on vision alone.

Multimodality is central to this approach, not as a simple aggregation of sensory outputs, but as an integrated interaction paradigm. Auditory feedback conveys descriptive and spatial information, haptic cues support directional and proximity awareness, and semantic interpretation enables higher-level understanding of objects, layouts, and interaction opportunities. To support this level of mediation, the architecture employs VLLMs as a central interpretative layer, enabling complex visual environments to be translated into adaptive, non-visual representations.

The remainder of this section details the design rationale underlying the architecture, outlining its design goals, inclusive principles, multimodal interaction strategy, and the role of Vision Large Language Models in supporting accessible immersive interaction.

4.1 Design Goals

The design goals of the proposed architecture were defined to address the limitations identified in existing accessibility solutions for immersive virtual environments, particularly their reliance on vision-centered interaction, application-specific adaptations, and low-level sensory feedback.

Autonomy is a primary design goal. Accessible VR interfaces should enable users to independently explore, orient themselves, and interact with virtual environments without requiring external assistance or prior familiarity with the space Lahav [2022]. For users with total blindness, this entails providing sufficient non-visual information to support the construction of accurate mental models of the environment. For users with low vision, autonomy requires mechanisms that supplement residual vision without overwhelming or conflicting with it. The architecture therefore aims to support self-directed navigation and decision-making through contin-

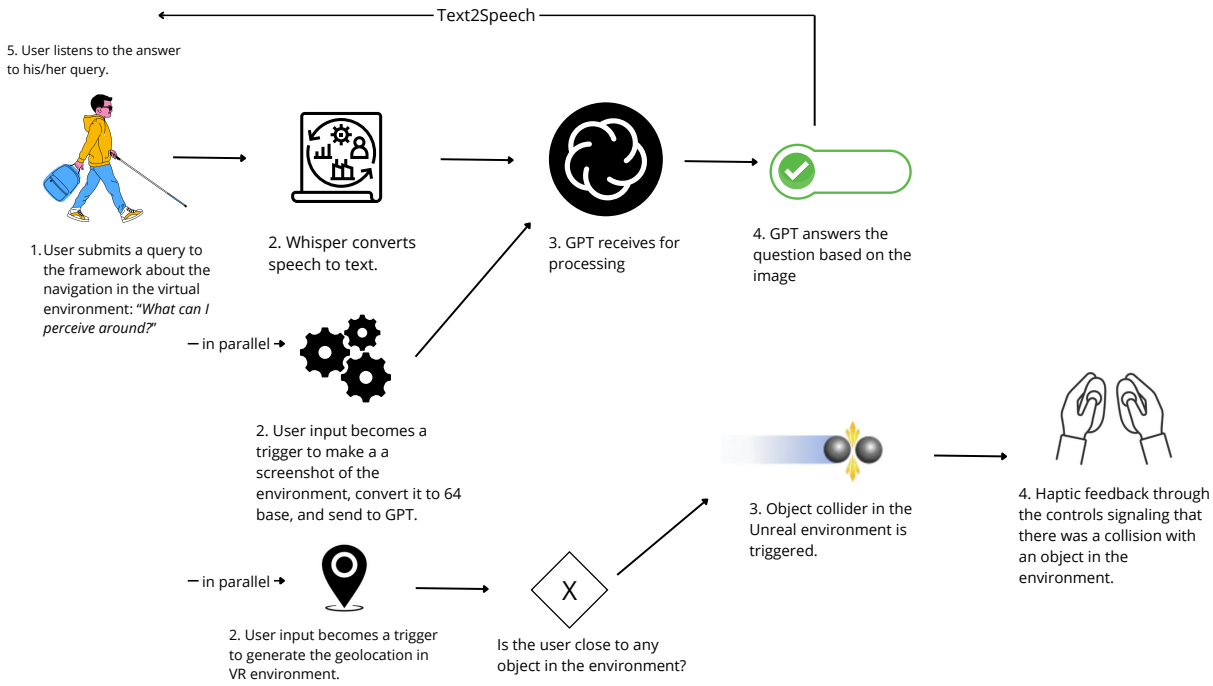

Figure 5. System architecture for AI-based interactive assistance integrated with Unreal Engine. The diagram illustrates the workflow of the proposed system. User input is processed by Whisper-Tiny for transcription and combined with a scene screenshot and spatial pose data (position and orientation) 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.

uous, context-aware feedback rather than scripted guidance or rigid interaction flows.

A second key goal is **low cognitive load**. Immersive environments are inherently information-rich, and the addition of accessibility layers risks introducing excessive or poorly structured feedback Raybourn *et al*. [2019]. Prior work has shown that visually impaired users—particularly those with low vision—are susceptible to cognitive overload when forced to integrate degraded visual input with dense auditory or haptic signals. To mitigate this risk, the architecture emphasizes selective, semantically grounded information delivery, prioritizing relevance over completeness. By leveraging AI-based scene interpretation, the interface aims to abstract complex visual data into concise, meaningful representations that reduce the mental effort required to interpret and act upon environmental cues.

Adaptability constitutes the second core design goal. Visual impairment manifests differently across individuals and contexts, and accessible interfaces must accommodate varying sensory preferences, residual abilities, and situational demands. The proposed architecture is designed to adapt both to different user profiles—such as total blindness versus low vision—and to different virtual environments with distinct spatial and interaction characteristics. Rather than encoding fixed mappings between environmental features and sensory outputs, the interface supports flexible modulation of feedback modalities, granularity, and semantic depth, enabling personalization and future extension.

Finally, the architecture prioritizes **generalizability and reusability** as design goals. Many existing accessibility solutions are tightly coupled to specific applications, limiting their broader impact. In contrast, the proposed interface is conceived as a domain-agnostic interaction layer that mediates access to virtual environments independently of their thematic content or functional purpose. By abstracting visual information into semantic representations that can be communicated through auditory and haptic channels, the architecture aims to support integration across diverse VR applications, including education, healthcare, training, and entertainment.

Together, these design goal form the foundation of the proposed architecture and guide subsequent design decisions related to multimodal interaction and AI-driven semantic mediation. In the following subsections, we discuss how inclusive design principles and multimodal AI techniques operationalize these goals within the proposed interface.

4.2 Inclusive Design Principles

Inclusive design is grounded in the recognition that human abilities, needs, and circumstances are inherently diverse and dynamic across the lifespan. Rather than assuming a "typical" user or projecting the designer's own capabilities onto others, inclusive design explicitly accounts for variability in age, ability, experience, and context. This perspective acknowledges both the temporal dimension of human life—where individuals may experience changing abilities over time—and the social dimension, in which people are interdependent and embedded in broader communities. As such, inclusive design seeks not only to address present needs but also to anticipate future conditions, supporting autonomy, participation, and quality of life over time.

From this theoretical standpoint, inclusive design emphasizes designing with users rather than for an abstract average user. Definitions proposed by institutions such as the Inclusive Design Research Centre describe inclusive design as an approach to creating products and environments that are us-

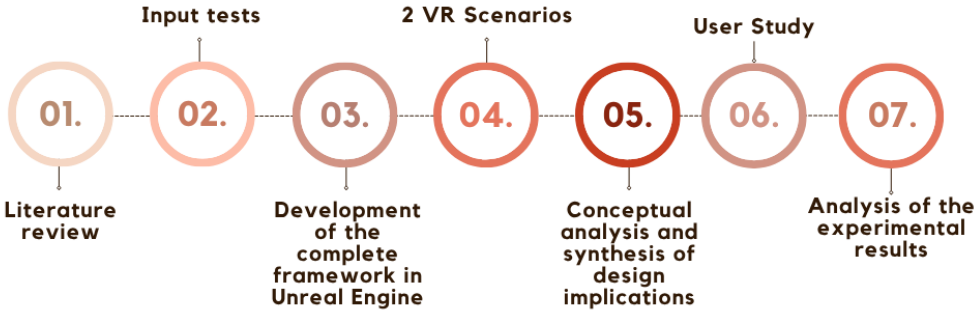

Figure 6. Research Method overview—seven sequential stages: 01) Literature review to map the state of the art and define research gaps; 02) Input tests to validate assumptions and tune data-capture/interaction parameters; 03) Architecture development in Unreal Engine integrating modules and logging; 04) Architecture instantiation in heterogeneous virtual scenarios to demonstrate architectural feasibility and generality; 05) Synthesis of architectural insights and design implications for accessible multimodal interfaces in virtual reality; 06) User study with participants with visual impairments to empirically validate the architecture across both immersive scenarios; 07) Analysis and synthesis of quantitative and qualitative results to derive empirical evidence and inform future design iterations.

able and appealing to the widest possible range of people, regardless of age, ability, or circumstance, by actively identifying and removing social, technical, and systemic barriers Pinheiro and da Silva [2010]. Closely related concepts—such as Universal Design and Design for All Story [2001]; Persson *et al*. [2015]—share this objective, differing primarily in terminology across regions, but converging on the principle that environments, products, and services should be understandable, usable, and accessible without requiring specialized adaptations or segregated solutions.

Within this tradition, inclusive design is understood as both a creative and ethical commitment. European architectures, including those articulated by the European Institute for Design and Disabilities and the Stockholm Declaration, frame Design for All as a holistic approach that integrates human diversity, social inclusion, and equality into planning, design, and governance processes. These architectures advocate for the involvement of users throughout the design process as an ideal toward which accessible systems should aspire. The proposed architecture is grounded in these mentioned inclusive design principles that treat human diversity as a fundamental design parameter rather than an edge case Persson *et al*. [2015]. In immersive virtual environments, this requires moving beyond assumptions of a "typical" fully sighted user and accommodating a broad range of sensory abilities, perceptual strategies, and interaction needs Dudley *et al*. [2023]. This perspective aligns with accessibility and human-centered design literature that advocates for systems usable by the widest possible audience without requiring separate or specialized adaptations. In the present study, user participation was incorporated at the evaluation stage through a user study with individuals with visual impairments, whose feedback informed the identification of design limitations and opportunities for refinement. The empirical findings reported in Section 6 contribute to grounding future iterations of the architecture in the lived experiences of the target population.

Equitable access to information is another guiding principle Dudley *et al*. [2023]; Rodrigues *et al*. [2025]. Beyond basic operability, inclusive VR design must enable access to meaningful semantic content, such as object identity, spatial relationships, and interaction affordances. The proposed architecture addresses this by translating visual content into semantically rich representations accessible through auditory and haptic modalities, allowing visually impaired users to engage with virtual environments at a comparable conceptual level.

Finally, the design prioritizes minimizing unnecessary complexity. Through semantic abstraction and context-aware mediation, the architecture presents only the most relevant information at a given moment, offloading perceptual and cognitive effort from the user to the system. Together, these principles ensure that accessibility is embedded as an integral aspect of the interface design, shaping how multimodal feedback and AI-based mediation support inclusive interaction in immersive virtual environments.

4.3 Why Multimodality (Audio, Haptics, and Semantics)

Multimodality constitutes a foundational design choice in the proposed architecture, reflecting the recognition that no single sensory channel is sufficient to support accessible interaction in complex immersive environments. VR systems are inherently spatial, dynamic, and information-dense, and relying on a single non-visual modality—such as audio or haptics alone—can impose significant cognitive and perceptual limitations Giudice [2018]. Consequently, the architecture adopts a multimodal interaction paradigm that integrates auditory feedback, haptic cues, and semantic interpretation to support robust, flexible, and inclusive access to virtual environments.

Auditory feedback plays a central role in conveying both spatial and descriptive information Wang and Ben-Arie [1996]. Through techniques such as spatialized audio and verbal descriptions, sound can communicate object locations, environmental boundaries, and interaction opportunities. However, audio-only solutions often struggle to balance informativeness with cognitive load, particularly in environments where multiple elements compete for attention. Excessive or poorly structured auditory output can overwhelm users, especially those with low vision who must simultaneously integrate residual visual input with auditory cues Brown and Proulx [2016].

Haptic feedback complements auditory information by

providing continuous, low-latency signals related to direction, proximity, and movement. Tactile cues are particularly effective for conveying spatial relationships and supporting motor coordination without demanding sustained cognitive attention Macaluso and Driver [2001]. In the proposed architecture, haptics are not treated as a substitute for audio descriptions, but as a parallel channel that reinforces spatial awareness and supports intuitive navigation. This complementary relationship allows users to distribute perceptual effort across modalities, reducing reliance on any single channel and mitigating sensory overload.

Crucially, the architecture extends beyond traditional multisensory designs by incorporating semantic mediation as an integral component of multimodality. While audio and haptic cues can effectively convey low-level spatial signals, they are insufficient for communicating higher-level meaning, such as object identity, functional relevance, or contextual relationships within the environment. Semantic information enables users to understand not only where things are, but also what they are and why they matter within a given context. By embedding semantic interpretation into the interaction loop, the architecture supports more informed decision-making and autonomous exploration.

The integration of audio, haptics, and semantics reflects a shift from reactive feedback toward proactive, context-aware interaction. Rather than responding solely to user actions or proximity events, the architecture continuously interprets the virtual environment and adapts multimodal feedback based on situational relevance. This design approach aligns with inclusive interaction principles by accommodating diverse perceptual strategies and enabling users to engage with immersive spaces at different levels of detail and abstraction.

4.4 Why Vision Large Language Models as the Mediation Layer

The selection of VLLMs as the core mediation layer of the proposed architecture is driven by the need to bridge low-level sensory data and high-level semantic understanding in immersive virtual environments. Traditional accessibility mechanisms in VR typically rely on pre-scripted mappings between visual events and sensory outputs, such as predefined audio cues or haptic signals Rahimi *et al*. [2025]. While effective in constrained scenarios, these approaches lack scalability and adaptability when applied to complex, dynamic environments. VLLMs offer a fundamentally different paradigm by enabling systems to interpret visual scenes Li *et al*. [2025a], reason about their content, and generate context-aware semantic representations in real time.

VLLMs integrate computer vision capabilities with natural language reasoning, allowing them to extract meaning from visual input and express that meaning in a form that is intelligible and actionable for users Liang *et al*. [2026]. In the context of accessibility, this capability is particularly valuable because it supports the translation of rich visual information—such as spatial layouts, object identities, and interaction affordances—into descriptions that can be delivered through non-visual modalities. This semantic mediation enables users to form coherent mental models of virtual environments, moving beyond reactive navigation toward informed, autonomous exploration.

From an inclusive design perspective, VLLMs support personalization and user control in ways that rule-based systems cannot. Because semantic descriptions are generated dynamically, the mediation layer can adapt its output to different user profiles, preferences, and levels of visual impairment. Users with total blindness may prioritize comprehensive environmental descriptions, while users with low vision may benefit from selective, disambiguating information that supplements residual vision. This adaptability avoids imposing a one-size-fits-all interaction model, which has been a recurring limitation of prior accessibility solutions. Importantly, positioning VLLMs as a central mediation layer shifts the role of AI from a peripheral assistive feature to a foundational component of the interaction architecture, enabling a tighter coupling between perception, reasoning, and interaction. The specific mechanisms through which this mediation operates within the interface architecture are detailed in Section 5.2.

5 Interface Architecture

This section presents the architecture of the proposed interface, describing how visual information from immersive virtual environments is semantically mediated and translated into multimodal feedback. The focus is on the interface structure and interaction flow, illustrating how the architecture operationalizes the design principles discussed earlier and supports integration across diverse VR applications.

The interface adopts a modular, pipeline-based architecture in which visual data from the virtual environment is interpreted by a Vision Large Language Model that functions as a semantic mediation layer. The resulting context-aware representations are mapped to auditory and haptic feedback, enabling non-visual perception of spatial layouts, objects, and interaction opportunities.

A key architectural feature is the decoupling of accessibility mechanisms from application-specific logic. By operating as an intermediate layer between the VR environment and the user, the interface supports reuse across domains and interaction contexts, reinforcing its generalizability and extensibility.

The remainder of this section details the conceptual pipeline and interaction flow, examines the role of the VLLM in semantic abstraction, and describes how user inputs are translated into multimodal feedback.

5.1 Conceptual Pipeline and Interaction Flow

The proposed interface is structured around a conceptual pipeline that governs the flow of information between the immersive virtual environment and the user, integrating perception, interpretation, and multimodal feedback into a unified interaction loop. This pipeline operationalizes the architecture's design goals by ensuring that user interaction is continuously mediated through semantic understanding rather than direct reliance on visual perception.

At the first stage of the pipeline, the virtual environment provides raw visual and spatial data, including scene geometry, object configurations, and dynamic changes resulting from user actions or environmental events. This information is captured independently of any application-specific logic, allowing the interface to operate as an intermediate layer be-

tween the VR system and the user.

The captured visual data is then processed by the Vision Large Language Model, which performs scene understanding and semantic abstraction. Rather than transmitting low-level visual features, the VLLM interprets the environment at a conceptual level, identifying relevant objects, spatial relationships, and interaction affordances. This semantic representation serves as the core internal state of the interface, enabling the system to reason about the environment in a manner that is both human-interpretable and adaptable across contexts.

User input constitutes the next component of the interaction flow and includes locomotion and controller actions supported by the VR platform, and proximity or collision events generated during navigation. These inputs are not treated as isolated commands but are interpreted in relation to the current semantic state of the environment.

Based on this combined interpretation of environmental state and user input, the interface generates multimodal output through auditory and haptic channels. Auditory feedback conveys descriptive and spatial information, while haptic cues communicate directional, proximity-based, or movement-related signals. Crucially, these outputs are derived from the semantic representation produced by the VLLM, ensuring that feedback remains meaningful, context-aware, and aligned with the user's goals rather than reacting solely to low-level events.

The interaction flow is continuous and iterative. As users move through the environment and interact with virtual elements, the pipeline updates the semantic representation and adapts multimodal feedback in real time. This closed-loop design supports dynamic exploration and reduces the cognitive effort required to interpret changes in the environment, reinforcing the architecture's emphasis on user autonomy.

5.2 Role of the Vision Large Language Model in Scene Understanding and Semantic Abstraction

Within the proposed interface architecture, the VLLM functions as the central interpretative component responsible for transforming raw visual input into structured, semantically meaningful representations. Rather than operating as a peripheral assistive module, the VLLM constitutes the core mediation layer that enables accessibility through abstraction, reasoning, and contextual interpretation.

At the level of scene understanding, the VLLM processes visual information originating from the virtual environment to identify salient elements such as objects, spatial layouts, relative positions, and interaction affordances. This process goes beyond pixel-level analysis or object detection by incorporating contextual reasoning about how elements relate to one another within the environment. For example, instead of recognizing isolated objects, the model can infer functional groupings, navigable paths, and regions of interest, which are critical for meaningful interaction in immersive spaces. To assess the precision of GPT-4o in interpreting rendered virtual scenes, we manually evaluated the model's responses against a set of 26 static screenshots—14 from the hospital reception hall and 12 from the classroom—submitted independently during the validation phase (Step 2). Each response was rated as correct, partially correct, or incorrect by comparing the

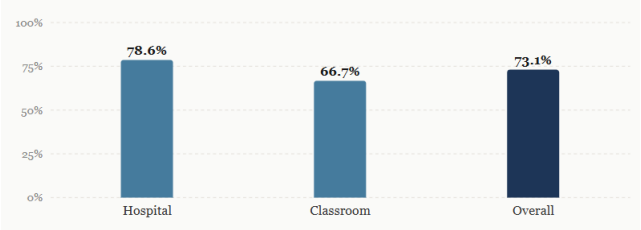

Figure 7. GPT-4o scene interpretation accuracy across 26 screenshots from the hospital and classroom virtual environments. Responses were manually evaluated against ground-truth scene descriptions and classified as correct or erroneous. Error categories include spatial configuration misidentifications and color misclassifications.

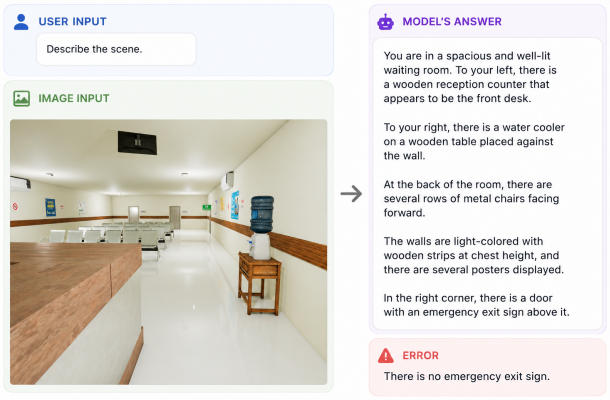

Figure 8. Representative GPT-4o output during the model validation phase (Step 2): a screenshot from the hospital reception hall submitted as visual input and the corresponding semantic description generated by the model.

model's output against the known ground truth of each scene. The evaluation yielded an overall accuracy of 73.1% (19 out of 26 images correctly described), with 78.6% accuracy in the hospital scenario and 66.7% in the classroom. Observed errors were primarily of two categories: spatial configuration misidentifications (e.g., confusing a U-shaped counter with an L-shaped one, or incorrectly reporting the number of chair columns or doors) and color misclassifications (e.g., reporting orange instead of red for a highlighted chair in the classroom environment). These error types, while relevant, were generally non-critical for navigation purposes, as they did not prevent participants from completing tasks during the user study. This result suggests that while GPT-4o provides reliable semantic mediation in most cases, further prompt refinement or post-processing strategies may be needed to address spatial and chromatic ambiguities in rendered virtual environments. These results are summarized in Figure 7, which presents the accuracy breakdown by scenario and error category. Figure 8 illustrates a representative model output, showing the scene image submitted as input alongside the corresponding textual description generated by GPT-4o.

The output of this process is a semantic abstraction of the environment—a representation that captures what is relevant for interaction rather than everything that is visually present. This abstraction acts as an intermediate layer between the immersive system and the user, enabling the interface to reason about the environment at a conceptual level. By prioritizing semantic relevance over visual fidelity, the architecture reduces informational complexity and supports accessibility by filtering out visually dense or redundant details that may otherwise overwhelm users with visual impairments.

Crucially, semantic abstraction enables the decoupling of accessibility logic from environment-specific implementations. Because the VLLM operates on high-level concepts rather than fixed visual features, the same mediation strategy can be applied across diverse virtual environments without reauthoring rules or mappings for each scenario. This property directly supports the architecture's design goals of generalizability and reusability, distinguishing it from rule-based or pre-scripted accessibility solutions. Another important function of the VLLM lies in its capacity for adaptive semantic reasoning. As users interact with the environment, the semantic representation is continuously updated to reflect changes in context, user position, and interaction state. This dynamic interpretation allows the interface to adjust the granularity and focus of feedback in real time, emphasizing information that is most relevant to the user's current goals and actions. Such adaptability is essential for supporting autonomous exploration in immersive environments. Finally, by expressing semantic representations in a language-compatible format, the VLLM facilitates seamless integration with multimodal feedback channels. Semantic content generated by the model can be translated into natural language descriptions for auditory output or mapped onto structured signals for haptic feedback. This capability enables coherent cross-modal communication and ensures that different sensory channels convey consistent and complementary information, reinforcing users' mental models of the virtual space.

5.2.1 VLLM Reliability and Mitigation Strategies

The integration of Vision Large Language Models as a core mediation layer introduces reliability challenges that must be explicitly addressed in the context of a dynamic, context-aware accessibility system. Three failure modes are of particular concern. **First**, overly general outputs occur when the model produces plausible but spatially unspecific descriptions that lack the directional or object-level detail necessary to support navigation decisions. In environments with high object density or complex layouts, such responses may leave users without actionable guidance. **Second**, hallucinations — instances in which the model confidently describes objects, spatial configurations, or paths that are not present in the rendered scene — pose a more critical risk, as they may actively mislead navigation rather than merely fail to assist it. **Third**, topic divergence arises when the model's response drifts away from the navigational intent of the user's query, producing contextually irrelevant content that increases cognitive load and disrupts interaction flow.

To mitigate these challenges, the architecture adopts a set of structural and prompt-level strategies. The system prompt is explicitly scoped to spatial and navigational content, constraining the model's response to information relevant to the user's current viewpoint and task. Visual grounding is enforced by capturing rendered frames at the moment of each query, ensuring that model responses are anchored to the user's actual perspective rather than a cached or outdated scene representation. Prompt refinement conducted during the validation phase (Step 2) was specifically aimed at reducing response vagueness and improving directional specificity, guided by the error categories identified in the 26-screenshot evaluation.

Figure 9. Scene of the developed classroom environment.

Acceptability of VLLM Accuracy for Navigation Tasks. While the observed accuracy of 73.1% indicates that errors occur in approximately one quarter of cases, it is important to contextualize this result in terms of functional impact rather than raw prediction correctness. First, the majority of observed errors were non-critical for navigation, consisting primarily of spatial imprecision (e.g., approximate layouts) and color misclassification. These inaccuracies did not prevent participants from successfully completing tasks during the user study, suggesting that perfect scene reconstruction is not a strict requirement for effective accessibility in this context. Second, the system does not rely on single-shot decisions but instead operates as an interactive, continuous feedback loop. Users receive updated semantic descriptions as they move and explore the environment, allowing them to correct potential misunderstandings through subsequent interactions. This incremental and redundant information flow reduces the risk associated with isolated model errors. Third, navigation is supported through multimodal redundancy. Auditory descriptions are complemented by proximity-triggered haptic feedback, which provides direct spatial cues independent of the VLLM's semantic interpretation. This design ensures that even in the presence of incomplete or imprecise semantic descriptions, users retain access to reliable low-level guidance for obstacle avoidance and target localization. Finally, the evaluation did not reveal instances in which VLLM errors led to unsafe or irrecoverable navigation outcomes. Instead, errors manifested as minor inefficiencies (e.g., slight detours or hesitation) rather than critical failures. This distinction is crucial: in accessibility contexts, the acceptable threshold for model performance is determined by its impact on task completion and user safety, rather than strict semantic accuracy.

5.3 Architecture Instantiation and Illustrative Scenarios

To illustrate how the proposed interface architecture can be instantiated in practice, we integrated it into immersive applications developed using Unreal Engine 5. This instantiation serves as a concrete demonstration of how the architecture operates alongside existing VR projects, without altering application-specific logic or interaction design. The goal of this subsection is not to evaluate performance or user outcomes, but to exemplify how the architectural components described earlier are realized within similar immersive environments.

Scenario	Model Output (Error)	Ground Truth / Impact
Classroom	"There is an orange chair aligned in front of the room."	Chair was red, not orange. Color misclassification did not affect task completion, as object identity and position were correctly conveyed.
Classroom	"There are three columns of chairs facing the board."	Only two columns were present. Spatial overestimation caused minor imprecision but did not prevent navigation.
Hospital Reception	"An L-shaped counter is positioned ahead of you."	The counter was U-shaped. Despite the incorrect geometry, the model correctly conveyed the presence and navigational relevance of the counter.
Hospital Reception	"There is a single row of seats on your left."	Four rows of seats were present. Underestimation of object quantity did not impact obstacle avoidance due to haptic feedback.

Table 1. Representative examples of VLLM errors observed during the validation phase, illustrating typical discrepancies between model outputs and ground truth. Most errors involve spatial approximation or color misclassification and were non-critical for task completion.

Figure 10. Scene of the developed hospital environment.

At the system level, the architecture operates by intercepting visual and spatial information generated by the Unreal Engine rendering pipeline and routing it through the semantic mediation layer. High-resolution rendered frames are captured and processed by a VLLM, which interprets the visual scene and produces structured semantic representations as shown in Figure 9 and Figure 10 . User input, such as voice commands, is transcribed via a Speech-to-Text module and incorporated into the interpretation process. The resulting semantic output is then mapped to auditory and haptic feedback, delivered through the engine's audio system and force-feedback mechanisms. This architecture enables the architecture to function as an accessibility layer that augments immersive applications without imposing constraints on their internal design.

To demonstrate the generality of the architecture across different spatial and interaction contexts, we instantiated it within two ecologically grounded but structurally distinct VR scenarios: a classroom and a hospital reception hall. These environments were selected not as experimental testbeds, but as illustrative use cases representing common categories of immersive applications. The classroom scenario exemplifies a structured environment with fixed object placement and localized interaction zones, typical of educational or training applications. In contrast, the hospital reception hall represents a more open and socially oriented space, characterized by broader navigation demands, multiple points of interest, and transitional areas such as corridors and service desks.

In both scenarios, the architecture provides continuous semantic mediation of the environment, enabling users to receive audio-based descriptions of spatial layout, object presence, and interaction opportunities. Haptic feedback is employed to convey directional guidance and proximity information, with intensity patterns dynamically modulated based on the user's spatial relationship to relevant elements in the scene. Importantly, these interaction mechanisms are generated by the architecture independently of the specific content or narrative of the application, reinforcing its role as a reusable accessibility architecture rather than a solution tailored to a particular scenario.

5.4 Reusable Components, Interfaces, and Extension Points

A key characteristic of the proposed system is its modular and reusable architecture, designed to support multiple scenarios without requiring structural changes. Rather than implementing isolated, scenario-specific solutions, the system is organized around a set of shared components that are consistently reused across environments. Evidence from both the Hospital and Classroom Blueprints, as well as the Python-based backend, indicates a high degree of structural regularity, suggesting that these environments are instantiations of a common architectural backbone rather than independent prototypes.

This design follows a clear separation of concerns: core functionalities are encapsulated into reusable modules with well-defined interfaces, while scenario-specific behavior is expressed through composition. As a result, the system achieves both flexibility and scalability, enabling the rapid development of new environments with minimal engineering overhead. Figure 11 illustrates the package diagram of the proposed system, highlighting its modular structure and component organization.

5.4.1 Core Reusable Modules

The architecture is grounded on a set of core modules that implement the system's fundamental capabilities. These components are intentionally designed to be scenario-agnostic, ensuring that their behavior remains consistent regardless of the environment in which they are deployed. The networking component (TCP client) abstracts communication with

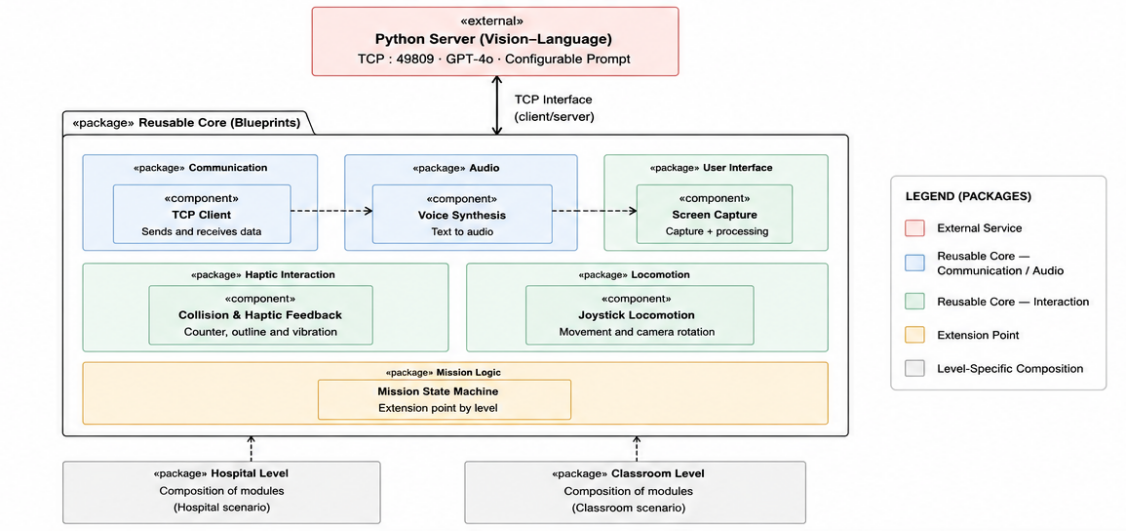

Figure 11. Package diagram of the proposed architecture, illustrating the reusable core modules, their interfaces, and the composition of scenario-specific levels (Hospital and Classroom) interacting with the external vision-language server.

the external Python server through a persistent socket interface. Its design is content-agnostic, supporting heterogeneous data types—including navigation commands, textual descriptions, and image payloads—over a unified channel. As such, the identical structure of this component across levels reinforces its role as a foundational infrastructure layer rather than scenario-specific logic. Building on this communication layer, the text-to-speech (TTS) module implements a unidirectional transformation pipeline that converts textual input into spatialized audio feedback. In this case, the voice provider is abstracted behind a stable interface, enabling the interchangeable use of external APIs (e.g., ElevenLabs or alternative services). This design choice reflects a deliberate decoupling between interface and implementation, allowing the audio synthesis backend to evolve independently. Similarly, the screen capture component encapsulates the acquisition and transmission of visual data to the backend. Its responsibilities are strictly limited to triggering image capture, attaching contextual metadata, and forwarding the result via the networking module. By doing so, and by avoiding embedded mission logic, the component remains fully reusable across tasks and environments. In parallel, the collision and haptic feedback module consolidates interaction-related behaviors into a unified abstraction. Although it supports distinct interaction types (e.g., objective-related feedback versus obstacle detection), all interactions rely on a shared event-driven mechanism. This common foundation ensures consistency while still allowing higher-level behavioral differentiation. Finally, the locomotion module isolates user movement from scenario logic by decoupling input handling from spatial navigation. Directional movement is computed relative to the camera frame, while rotational control is handled independently, thereby ensuring that locomotion behavior remains reusable and unaffected by mission-specific constraints.

5.4.2 Scenario-Specific Composition

While the core modules remain invariant, scenario variability is introduced through composition rather than modification. This principle enables the system to support diverse environments without altering its underlying structure. At the center of this process is the mission state machine, which acts as the primary coordination layer. It binds reusable components to scenario-specific objectives by defining sequences of interactions, associating feedback with mission states, and providing contextual information to the backend.

Consequently, introducing a new scenario does not require modifying existing modules. Instead, it involves configuring a new set of states, defining spatial triggers (e.g., colliders), and composing existing components accordingly. The Hospital and Classroom environments can therefore be interpreted as different configurations of the same architectural architecture.

5.4.3 Extension Points

Beyond reuse, the architecture is explicitly designed to support extensibility, allowing new capabilities to be integrated without disrupting existing components. Several extension points are exposed. The TTS provider is treated as a replaceable dependency, enabling seamless integration of alternative speech synthesis services. Similarly, the vision-language processing pipeline is encapsulated within the backend, allowing model selection and prompt design to evolve independently of the client-side system. In addition, new scenarios can be introduced solely through the definition of mission states and spatial configurations, without requiring changes to reusable modules. The input system also remains extensible, supporting the incorporation of new interaction modalities without impacting locomotion or mission logic.

Together, these design choices ensure that the system is not only reusable but also adaptable, providing a robust foundation for future research and development.

6 Evaluation

To investigate the reliability of the proposed architecture, we conducted a case study, an exploratory empirical method suited to investigating a phenomenon in its natural setting using data gathered from a small number of entities (in this case, participants with visual impairments interacting with the system). This method is particularly appropriate for our evaluation, as it addresses a contemporary phenomenon – AI-

mediated multimodal interaction in immersive environments – observed in a realistic, ecologically grounded context, and it draws on multiple sources of evidence, including quantitative task metrics, structured questionnaire responses, and qualitative interview data.

The case study was conducted according to the following steps Runeson and Höst [2009]: (i) case study design, encompassing the preparation and planning of data collection procedures; (ii) execution, in which evidence is systematically collected through user interaction with the system; (iii) analysis of the collected data; and (iv) reporting of the findings.

In the **design** stage, we defined the target population (individuals with total blindness and low vision, recruited through a partnership with a local association for the blind), the two ecologically valid scenarios (a classroom and a hospital reception hall) representing structurally distinct interaction contexts, the sequence of goal-oriented tasks to be completed exclusively through auditory and haptic feedback, and the quantitative (task completion time, number of collisions) and subjective (Likert-scale questionnaire, semi-structured interview) measures to be collected, as detailed in Sections 6.1 and 6.3. This stage was conducted under an approved institutional ethics protocol (Protocol 88442725.6.0000.5083).

In the **execution** stage, eight participants were recruited, of whom five completed the full protocol (N=5); each session followed a fixed procedure comprising informed consent, a demographic questionnaire, headset fitting and calibration, a practice scenario, and task execution across both immersive environments in a fixed order, as described in Section 6.2.

In the **analysis** stage, quantitative data were summarized descriptively by participant group (blind and low vision) and task, while subjective questionnaire data were analyzed at the item level (mean and standard deviation across the nine evaluated dimensions). Open-ended responses and interview transcripts were examined through thematic analysis following a top-down approach Braun and Clarke [2006], guided by the study's research questions.

Finally, in the **reporting** stage, the results of this analysis are presented in Section 6.4 and further interpreted in light of the architecture's design goals and limitations in Section 7.

6.1 Participants

Eight participants with visual impairments were initially recruited through partnerships with local organizations and community networks (e.g., Association of the Blinds of the State of Goiás - Brazil), of whom five successfully completed the full evaluation protocol (N=5). Three participants were unable to complete the study due to difficulties unrelated to the architecture itself: two experienced significant challenges in operating the VR controller scheme — specifically in memorizing and executing the distinct button mappings required for locomotion, camera rotation, and AI interaction — while one participant, an elderly individual with limited literacy, encountered compounding difficulties in following verbal instructions and managing the physical demands of the headset. These cases were documented and are discussed as design implications in Section 7.

The five participants who completed the study comprised three individuals with total blindness and two individuals with low vision due to cataracts and glaucoma. Although the sample size is modest, it is consistent with established practices in accessibility and assistive technology research involving users with disabilities, where recruitment is inherently constrained by the specialized nature of the population and the ethical demands of inclusive research design. Moreover, prior work Kudo *et al*. [2022]; Mack *et al*. [2021] has demonstrated that small, well-characterized samples can yield meaningful and generalizable insights, particularly in exploratory evaluations of novel interaction paradigms. In such contexts, the depth of perceptual evaluations, consistency of observed behaviors, and convergence of user experiences are often more critical. Our study follows this paradigm by prioritizing rich interaction data, detailed observation, and iterative validation of the proposed system. Importantly, this work should be understood as a foundational step toward larger-scale evaluations. The primary goal is to validate the feasibility and effectiveness of the proposed VLLM-mediated interaction architecture, providing empirical evidence and methodological grounding for future studies with expanded and more diverse participant groups.

The sample comprised four female and one male participants, aged between 24 and 78 years, representing different life stages and varying levels of prior experience with assistive technologies. Participants with total blindness relied entirely on the non-visual feedback provided by the system, while low vision participants were able to supplement this feedback with residual visual perception. Informed consent was obtained from all participants prior to their involvement.

6.2 Procedure

Upon arrival, participants were informed about the study objectives and signed an informed consent form. All data were treated confidentially and anonymously in accordance with the ethical guidelines approved by the Research Ethics Committee. Participants completed a short demographic questionnaire covering assistive technology usage and prior VR experience. Each participant was then fitted with a Meta Quest 3 headset, adjusted for individual comfort, and equipped with both Meta Quest Touch Plus Controllers (left and right units) — controllers that provide three degrees of freedom for hand tracking and include haptic vibration motors, used to deliver the proximity-based and collision force-feedback described in the architecture's haptic interaction flow — with wrist straps secured prior to each session. Basic calibration of orientation was performed when necessary. A brief training session in a dedicated practice scenario was conducted to familiarize participants with the VR equipment and the interaction mechanics of the architecture prior to the experimental tasks. Following the practice session, participants were immersed in both virtual scenarios—the classroom and the hospital reception hall—in a fixed order.

In the classroom scenario, participants were assigned two sequential tasks:

- Locate a highlighted chair and perform a sitting interaction;
- Navigate back to the entrance door and exit the room.

In the hospital scenario, participants were assigned three sequential tasks:

- Approach the reception desk and interact with the attendant;
- Navigate from the desk to the bathroom;
- Return from the bathroom to the main exit.

All navigation and interaction were guided exclusively by the multimodal feedback generated by the architecture, with no visual assistance provided. For participants with total blindness, the provided feedback constituted the sole available perceptual channel. In this context, a no-assistance condition is neither methodologically meaningful nor ethically appropriate, as it would remove the only viable means of perceiving the environment and reduce the task to a guaranteed-failure scenario. This consideration is consistent with established practices in accessibility research involving blind users Maidenbaum *et al*. [2016]. For participants with low vision, in contrast, the multimodal feedback complemented—rather than replaced—their residual visual perception. Although this group also had access to the VLLM-based semantic feedback and direct haptic guidance (i.e., without relying on the baseline condition), we observed that they made minimal practical use of the AI assistance. In most cases, their residual vision was sufficient to complete the tasks independently.

The primary effect observed was a slight increase in completion time during the initial tasks, typically on the order of a few seconds. This delay was not associated with difficulties in locating relevant elements in the environment, but rather with the additional cognitive effort of attempting to read automatically generated captions while simultaneously processing the audio descriptions. As participants became familiar with the system, this behavior diminished, and performance converged to that of the baseline condition. Importantly, we observed that the contribution of the AI was not uniform across tasks. Its utility became more evident in scenarios that required identifying specific semantic targets, such as locating "exit" or "restroom" signs to complete mission objectives. In these cases, the VLLM-based semantic interpretation provided clear benefits by enabling access to information that was not easily discernible through residual vision alone. In contrast, for tasks involving general navigation or object localization, participants were often able to rely on their residual vision to successfully complete the missions without substantial assistance from the AI. This suggests that, in low vision contexts, the architecture acts primarily as a complementary perceptual layer, becoming particularly valuable in situations that demand semantic understanding rather than purely spatial perception. To assess the specific contribution of VLLM-based semantic mediation in a setting where an independent perceptual channel exists, we conducted an additional within-subject comparison with the two low vision participants. For this subgroup, residual vision provides a non-zero baseline, enabling a with/without comparison Qiu *et al*. [2024]. This constraint is discussed in Section 7.

Short breaks were permitted between scenarios as needed. At the end of the session, participants completed a structured questionnaire and took part in a brief semi-structured interview.

6.3 Measures

Objective measures logged during task execution included task completion time in seconds and number of collisions. Subjective measures were collected via a structured questionnaire using a 5-point Likert scale, combining items adapted from the System Usability Scale (SUS) Brooke *et al*. [1996] with items developed specifically for this study. The questionnaire covered nine dimensions: usability and interaction design, accessibility and inclusivity, emotional and cognitive impact, functionality and performance, multisensory experience and engagement, presence and immersion, audio and narration, haptic feedback, and navigation and orientation. Another post-task subjective data were collected through open-ended questions addressing perceived facilitators, barriers, suggestions for improvement, and potential real-world applications. Responses were analyzed using thematic analysis following a top-down approach Braun and Clarke [2006], guided by the research questions.

6.4 Results

6.4.1 Task Performance

All five participants successfully completed the full sequence of tasks in both scenarios using exclusively auditory and haptic feedback. Table 2 summarizes mean task completion times and average number of collisions per group across the five tasks.

Participants with total blindness required longer completion times across all tasks compared to those with low vision, which was expected given their exclusive reliance on non-visual feedback. However, performance differences between groups diminished across tasks, suggesting a learning effect as participants became more familiar with the multimodal feedback patterns. Haptic feedback proved particularly effective for collision avoidance, with low vision participants averaging fewer collisions across all tasks.

6.4.2 Questionnaire Results

Subjective ratings indicated overall positive experiences across most dimensions. Usability and interaction design received high ratings (M=4.32, SD=0.18), with participants reporting that controls were learnable and tasks completable without external assistance. Emotional and cognitive impact was rated at ceiling (M=5.00, SD=0.00), reflecting consistent reports of confidence, motivation, and enjoyment. Functionality and performance (M=4.78, SD=0.09) and multisensory experience and engagement (M=4.76, SD=0.06) also received high scores, indicating that auditory descriptions and haptic cues were perceived as reliable and effective. Presence and immersion ratings were similarly strong, with participants reporting a consistent sense of being inside the environment and coherent system responsiveness. The dimension of accessibility and inclusivity received the lowest mean rating (M=2.85, SD=1.04), driven primarily by low scores on visual legibility items (L1–L3, M=2.60, SD=2.19), which were most relevant to low vision participants. Navigation and orientation showed internal variability: while landmark-based guidance received maximum ratings (N1, M=5.00), formation of a spatial mental map proved challenging (N4, M=1.60, SD=0.55), and interaction controls presented difficulties particularly for participants with limited prior technology experience (IC1, M=2.80, SD=0.84). The complete item-level questionnaire results are presented in Table 3.

Table 2. Task performance metrics across groups and tasks. Values are means. T1: sit on a chair (classroom); T2: exit the classroom; T3: approach the reception desk (hospital); T4: navigate to the bathroom (hospital); T5: exit the hospital.

Group	T1 (s)	T2 (s)	T3 (s)	T4 (s)	T5 (s)
Completion time (mean)					
Blind (n=3)	300	270	216	252	300
Low Vision (n=2)	120	144	150	132	114
Total (n=5)	228	220	190	204	226
Average collisions					
Blind	5	6	6	9	4
Low Vision	2	2	2	4	3

6.4.3 Questionnaire-based Findings

Thematic analysis of open-ended responses and interview transcripts yielded four recurring themes. **Navigation and spatial orientation** emerged as the most prominent theme, with participants acknowledging the effectiveness of auditory descriptions and haptic cues in supporting safe traversal, while also noting that more granular directional information—particularly during orientation changes—would be beneficial. **Task engagement and confidence** was consistently reported as high, with participants expressing that the system reduced uncertainty and supported a sense of autonomous control throughout the tasks. **Multisensory feedback complementarity** was identified as a key facilitator, with participants valuing the coordination between auditory descriptions and haptic vibrations as mutually reinforcing channels for spatial awareness. Finally, **areas for improvement** centered on control scheme complexity—particularly for older participants—and the need for more adaptive audio verbosity in environments with higher obstacle density.

Representative participant statements illustrating these themes include:

> "The descriptions from the system were helpful, but I would add even more detail about movements and directions, especially for navigating obstacles."

> "I felt I could perform the tasks without hesitation because the system guided me well."

> "The vibrations on the controller made it clear when I bumped into something, which helped me understand the space better."

Areas for improvement. Some participants suggested enhancements, particularly for more challenging or dynamic scenarios:

> "In some moments, the audio cues could be more detailed, for example, indicating distance or movement direction more clearly. That would help in environments with more obstacles."

7 Discussions

This work proposes and empirically evaluates an architectural approach to accessibility in immersive virtual environments, introducing a reusable architecture that mediates interaction through semantic, multimodal feedback. By positioning accessibility as a decoupled interface layer rather than an application-specific adaptation, the architecture addresses a key limitation of prior VR accessibility solutions: their tight coupling to individual scenarios, tasks, or user groups. This architectural decoupling enables scalable integration across different Unreal Engine applications and supports long-term maintainability of accessibility features.

Building on an earlier proof-of-concept, this paper advances the contribution by reframing the approach as a domain-agnostic interface architecture and providing empirical evidence of its feasibility through a user study with individuals with total blindness and low vision. The high ratings for usability, emotional impact, functionality, and multisensory engagement support the central claim that VLLM-based semantic mediation can enable meaningful non-visual navigation in immersive environments. The convergence of quantitative measures and participant feedback further suggests that the combination of auditory descriptions and proximity-triggered haptic feedback constitutes an effective and acceptable accessibility layer. However, our findings also reveal important differences across levels of visual impairment. While the architecture proved highly effective for participants with total blindness—serving as their primary perceptual channel—its impact for low vision (LV) participants was more limited. In practice, LV participants relied predominantly on their residual vision to complete tasks and made minimal use of the AI-mediated feedback. The main observable effect was a small increase in completion time during early interactions, associated with the additional cognitive effort of processing both visual and auditory information, rather than difficulties in navigation or task execution. These results suggest that the benefits of semantic mediation are not uniform across the visual impairment spectrum and are strongly dependent on task characteristics. In particular, the current task design did not heavily emphasize challenges where LV users typically experience greater difficulty, such as reading signage, interpreting fine-grained visual details, or disambiguating low-contrast elements. Our observations indicate that such scenarios are more likely to reveal the added value of VLLM-based semantic interpretation for this group. This highlights an important direction for future work: designing evaluation tasks that better capture the specific accessibility gaps experienced by LV users. The use of multimodal feedback grounded in semantic interpretation extends beyond traditional non-visual interaction techniques that rely on low-level cues or pre-scripted mappings. By translating visual content into context-aware auditory and haptic representations, the architecture supports meaningful navigation and interaction, particularly for users

Table 3. Questionnaire results per item (N=5). SD = sample standard deviation.

Item	Question	Mean	SD
U1	I learned how to interact quickly in the environment.	5.00	0.00
U2	The instructions were clear.	4.80	0.45
U3	Tasks could be completed without external help.	5.00	0.00
U4	Navigation/menus were simple to use.	4.40	0.55
U5	I felt confused by the controls.	2.40	0.55
P1	I felt like I was inside the environment.	5.00	0.00
P2	My attention stayed focused on the experience.	5.00	0.00
P3	The environment responded coherently to my actions.	5.00	0.00
P4	Interacting with virtual objects/people felt natural.	4.20	0.45
E1	The experience made me feel motivated.	5.00	0.00
E2	I felt confident while performing the tasks.	5.00	0.00
E3	Overall, the experience was enjoyable.	5.00	0.00
A1	Voice instructions were clear and easy to understand.	5.00	0.00
A2	Spatial audio (left/right/front) helped orientation.	5.00	0.00
A3	Audio volume was sufficient and consistent.	4.60	0.89
A4	Audio alerts (proximity/goal/error) were distinguishable.	3.80	0.45
L1	Contrast between text/icons and background was adequate.	2.60	2.19
L2	Font/icon size was sufficient for reading.	2.60	2.19
L3	Visual markers (arrows, highlights) helped notice useful objects.	2.60	2.19
HF1	Redundant signals (visual + audio + haptic) aided understanding.	5.00	0.00
HF2	Vibration/haptics helped confirm actions (grab, click, collide).	5.00	0.00
N1	Landmarks (beacons, continuous sounds) helped go from $A \rightarrow B$.	5.00	0.00
N2	The virtual guide/assistant helped me stay on track.	3.00	0.00
N3	I did not have unexpected collisions with objects/obstacles.	2.80	0.45
N4	A mental map of the scene became clear after a few seconds.	1.60	0.55
IC1	Gestures/buttons were easy to perform and remember.	2.80	0.84
IC2	Click/grab zones were tolerant to imprecision.	4.60	0.55
OB1	The initial tutorial covered the essentials to get started.	5.00	0.00
OB2	Replays/reminders of instructions were available on demand.	5.00	0.00
OB3	The experience pace respected my reading/listening speed.	5.00	0.00

with total blindness. Importantly, VLLMs are treated as a core mediation layer rather than auxiliary assistive tools, enabling adaptive and context-sensitive accessibility mechanisms. The validation of GPT-4o across 26 scene screenshots yielded an overall accuracy of 73.1%, with errors concentrated in spatial configuration and color classification — categories that, while relevant, proved non-critical for task completion in the evaluated scenarios.

Limitations. Despite these contributions, several limitations must be acknowledged. First, the user study involved five participants who completed the protocol, with three additional recruited individuals unable to complete the evaluation due to difficulties with the VR controller scheme and, in one case, limited literacy and prior technology exposure. While the sample size is consistent with established practices in accessibility research with specialized populations Whicher *et al*. [2018]; Picot *et al*. [2001]; Kudo *et al*. [2022], the findings should be interpreted as feasibility evidence rather than generalizable performance benchmarks. This limitation is particularly relevant for the LV subgroup (n=2), where even the inclusion of baseline comparisons would not support statistically meaningful conclusions. Future evaluations should expand the participant pool to include a broader range of visual impairment profiles, age groups, and technology literacy levels. Second, the architecture's effectiveness depends on the reliability of the underlying VLLM. The 73.1% scene interpretation accuracy observed during validation indicates that residual failure cases — including overly general outputs, spatial misidentifications, and potential hallucinations — remain a concern in dynamic deployment. Prompt engineering refinements, informed by the error categories identified in this study, represent a near-term avenue for improvement, but future work should investigate post-processing validation layers, confidence-based filtering, and retrieval-augmented strategies to improve output reliability. Third, cognitive load was not formally measured in the current evaluation. Although questionnaire results suggest that participants experienced low frustration and high confidence, the observed behavior in LV participants—particularly the initial delay caused by attempting to process multimodal information simultaneously—indicates that cognitive load may play a non-trivial role. Future studies should incorporate standardized workload instruments such as the NASA-TLX to provide objective evidence for cognitive accessibility claims,

as well as to better understand modality integration effects across user groups. Similarly, long-term habituation effects and performance across repeated sessions remain unexplored. Fourth, the current instantiation assumes the availability of standard VR input and output devices, specifically the Meta Quest 3 and its controllers. The difficulties encountered by three participants in operating the controller scheme highlight the need for simplified or adaptive input mechanisms. Future iterations should explore voice-only interaction modes, gesture-based alternatives, and controller configurations with reduced button complexity to broaden accessibility at the hardware level. Fifth, while user participation was incorporated at the evaluation stage through the formal user study, the architecture was not developed through a participatory design process involving users with visual impairments from its earliest conceptual stages. This represents a limitation with respect to inclusive design principles, which emphasize active user involvement throughout all phases of development. Future iterations should adopt more participatory approaches, engaging individuals with visual impairments as co-designers in defining interaction strategies, feedback modalities, and task structures. Finally, while the evaluated scenarios—a structured classroom and an open hospital reception hall—were selected to represent heterogeneous spatial contexts, future work should extend the architecture's evaluation to more dynamic and complex environments, as well as to tasks that more explicitly stress visual limitations in low vision populations, such as reading, object recognition under low contrast, and navigation in visually cluttered settings.

8 Conclusion

This paper presented a conceptual architecture for accessible interaction in immersive virtual environments, centered on semantic mediation through multimodal artificial intelligence. By integrating Vision Large Language Models with audio and haptic feedback within a reusable interface architecture, the proposed approach reframes accessibility as an architectural concern rather than an application-specific adaptation. This approach supports users with visual impairment, enabling meaningful navigation and interaction without assuming visual primacy.

The primary contribution of this work lies in the design of a layered interface architecture that demonstrates how semantic abstraction can bridge the gap between visually rich immersive environments and non-visual interaction modalities. Through its instantiation in heterogeneous scenarios, the architecture shows that accessibility mechanisms can be effectively decoupled from application logic and reused across diverse VR contexts. In addition, this work provides an experimental contribution in the form of a reproducible evaluation setup and empirical findings with visually impaired users. Beyond validating the proposed architecture, the study establishes a baseline that can be reused in future research, enabling replication, comparative analysis, and extension to new environments, models, or interaction strategies.

By emphasizing extensibility, generalizability, and inclusive design principles, this work contributes to ongoing discussions on how immersive technologies can be made more equitable and sustainable. Future research will focus on empirical evaluation, personalization strategies, and the integration of emerging multimodal AI models, further advancing the development of accessible, intelligent, and inclusive virtual reality systems.

Declarations

Acknowledgements

This work has been 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 of the MCTI grant number 057/2023, signed with EMBRAPII.

This manuscript is an extended and revised version of a previously published paper presented at the SVR 2025 conference, entitled *"A Multisensory AI-based architecture for Accessible Navigation of Visually Impaired Users in Virtual Environments: Preliminary Results."* The title and abstract of the present work have been substantially modified, and the content has been expanded with additional analyses, refinements, and discussions beyond the original publication.

Authors' Contributions

LC is the primary author and led the conception and writing of the manuscript. EA was responsible for investigation activities and participated in software development. GW and AB were involved in the implementation and development of the software components. VV, SL, and RS participated in manuscript review and editing. AG was responsible for funding acquisition. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The code files generated and/or analysed during the current study are available in Zenodo: `https://zenodo.org/records/18261634`, accessed on 02 September 2026.

Another supplementary material developed was a video demonstrating the execution of the application in one of the developed scenarios, and it is available on Dropbox: `https://www.dropbox.com/scl/fo/wvbyqyv57bkwbghpehiw6/ANUqZ-lUZKjuwdWc5c92ZuU?rlkey=4b7tvtidlxb0edfdtyvgou8zk&st=5a70u4jw&dl=0`, accessed on 02 September 2026.

Further relevant information

Use of Generative AI Tools: Generative artificial intelligence tools were used to assist in the drafting and refinement of the manuscript text. The AI support was limited to improving language clarity, organization, and readability, without generating original scientific contributions, experimental results, or analytical interpretations. All content was critically reviewed, validated, and finalized by the authors, who remain fully responsible for the accuracy, originality, and integrity of the work.

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. DOI: https://doi.org/10.3390/asi7030050.

Bendaly Hlaoui, Y., Zouhaier, L., and Ben Ayed, L. (2019). Model driven approach for adapting user interfaces to the context of accessibility: case of visually impaired users. *Journal on Multimodal User Interfaces*, 13(4):293–320. DOI: https://doi.org/10.1007/s12193-018-0277-z.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101. DOI: https://doi.org/10.1191/1478088706qp063oa.

Brooke, J. *et al*. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Brown, D. J. and Proulx, M. J. (2016). Audio–vision substitution for blind individuals: Addressing human information processing capacity limitations. *IEEE Journal of Selected Topics in Signal Processing*, 10(5):924–931. DOI: https://doi.org/10.1109/JSTSP.2016.2543678.

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. DOI: https://doi.org/10.1007/s10209-023-00969-0.

Delaney, B. and Biocca, F. (1995). Immersive virtual reality technology. *Communication in the age of virtual reality*, pages 57–124.

Dudley, J., Yin, L., Garaj, V., and Kristensson, P. O. (2023). Inclusive immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse. *Virtual Reality*, 27(4):2989–3020. DOI: https://doi.org/10.1007/s10055-023-00850-8.

Fernandes, H., Costa, P., Filipe, V., Paredes, H., and Barroso, J. (2019). A review of assistive spatial orientation and navigation technologies for the visually impaired. *Universal Access in the Information Society*, 18(1):155–168. DOI: https://doi.org/10.1007/s10209-017-0570-8.

Fernandes, P. C., Rocha, G. H. M., Adachi, B. H., Silvas, F. P., Coelho, B. N., and Delabrida, S. (2024). Exploring human interaction in virtual reality: An experience report on users with and without visual impairment. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*, pages 1–11. DOI: https://doi.org/10.1145/3702038.3702047.

Freire, A. P., Tavares, D. C., and Aranha, R. V. (2023). Pesquisa em interação humano-computador com pessoas com deficiência. In *Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC)*, pages 13–14. SBC. DOI: https://doi.org/10.5753/ihc_estendido.2023.233768.

Fryer, L. (2013). *Putting it into words: The impact of visual impairment on perception, experience and presence*. PhD thesis, Goldsmiths, University of London.

Giudice, N. A. (2018). Navigating without vision: Principles of blind spatial cognition. In *Handbook of behavioral and cognitive geography*, pages 260–288. Edward Elgar Publishing. DOI: https://doi.org/10.4337/9781784717544.00024.

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. DOI: https://doi.org/10.3389/frvir.2021.697504.

Kane, S. K., Morris, M. R., Perkins, A. Z., Wigdor, D., Ladner, R. E., and Wobbrock, J. O. (2011). Access overlays: improving non-visual access to large touch screens for blind users. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 273–282. DOI: https://doi.org/10.1145/2047196.2047232.

Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X., and Han, Y. (2025). Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36. DOI: https://doi.org/10.1145/3711680.

Kudo, T. N., de Freitas Bulcão-Neto, R., Neto, V. V. G., and Vincenzi, A. M. R. (2022). Aligning requirements and testing through metamodeling and patterns: design and evaluation. *Requirements Engineering*, 28(1):97. DOI: https://doi.org/10.1007/s00766-022-00377-5.

Lahav, O. (2022). Virtual reality systems as an orientation aid for people who are blind to acquire new spatial information. *Sensors*, 22(4):1307. DOI: https://doi.org/10.3390/s22041307.

Leat, S. J., Legge, G. E., and Bullimore, M. A. (1999). What is low vision? a re-evaluation of definitions. *Optometry and Vision Science*, 76(4):198–211.

Li, Y., Lai, Z., Bao, W., Tan, Z., Dao, A., Sui, K., Shen, J., Liu, D., Liu, H., and Kong, Y. (2025a). Visual large language models for generalized and specialized applications. *arXiv preprint arXiv:2501.02765*. DOI: https://doi.org/10.48550/arXiv.2501.02765.

Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., and Shi, G. (2025b). 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. DOI: https://doi.org/10.1109/CVPRW67362.2025.00147.

Liang, C. X., Tian, P., Yin, C. H., Yua, Y., Wei, A.-H., Li, M., Song, X., Wang, T., Bi, Z., Liu, M., *et al*. (2026). A comprehensive survey and guide to multimodal large language models in vision–language tasks. *Computation*, 14(6):125. DOI: https://doi.org/10.3390/computation14060125.

Liu, M., Chen, C., and Gurari, D. (2023). An evaluation of gpt-4v and gemini in online vqa. *arXiv preprint arXiv:2312.10637*. DOI: https://doi.org/10.48550/arXiv.2312.10637.

Liu, T., Fazli, P., and Jeong, H. (2024). Artificial intelligence in virtual reality for blind and low vision individuals: Literature review. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 1333–1338. SAGE Publications Sage CA: Los Angeles, CA. DOI: https://doi.org/10.1177/10711813241266832.

Macaluso, E. and Driver, J. (2001). Spatial attention and crossmodal interactions between vision and touch. *Neuropsychologia*, 39(12):1304–1316. DOI: https://doi.org/10.1016/S0028-3932(01)00119-1.

Mack, K., McDonnell, E., Jain, D., Lu Wang, L., E. Froehlich, J., and Findlater, L. (2021). What do we mean by "accessibility research"? a literature survey of accessibility papers in chi and assets from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18. DOI: https://doi.org/10.1145/3411764.3445412.

Maidenbaum, S., Buchs, G., Abboud, S., Lavi-Rotbain, O., and Amedi, A. (2016). Perception of graphical virtual environments by blind users via sensory substitution. *PloS one*, 11(2):e0147501. DOI: https://doi.org/10.1371/journal.pone.0147501.

Masnadi, S., Williamson, B., González, A. N. V., and LaViola, J. J. (2020). Vriassist: An eye-tracked virtual reality low vision assistance tool. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, pages 808–809. IEEE. DOI: https://doi.org/10.1109/VRW50115.2020.00255.

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. DOI: https://doi.org/10.1109/ISMAR-Adjunct.2019.00122.

Oliveira, A. D., dos Santos, P. S., Júnior, W. E. M., Aljedaani, W., Eler, D. M., and Eler, M. M. (2023). Análise de avaliaçoes de acessibilidade digital associadas a deficiências visuais e condiçoes oculares. In *Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC)*, pages 241–245. SBC. DOI: https://doi.org/10.5753/ihc_estendido.2023.233060.

Ortiz-Escobar, L. M., Chavarria, M. A., Schönenberger, K., Hurst, S., Stein, M. A., Mugeere, A., and Rivas Velarde, M. (2023). Assessing the implementation of user-centred design standards on assistive technology for persons with visual impairments: a systematic review. *Frontiers in Rehabilitation Sciences*, 4:1238158. DOI: https://doi.org/10.3389/fresc.2023.1238158.

Persson, H., Åhman, H., Yngling, A. A., and Gulliksen, J. (2015). Universal design, inclusive design, accessible design, design for all: different concepts—one goal? on the concept of accessibility—historical, methodological and philosophical aspects. *Universal access in the information society*, 14(4):505–526. DOI: https://doi.org/10.1007/s10209-014-0358-z.

Picot, S. J. F., Samonte, J., Tierney, J. A., Connor, J., and Powel, L. L. (2001). Effective sampling of rare population elements: Black female caregivers and non-caregivers. *Research on Aging*, 23(6):694–712. DOI: https://doi.org/10.1177/0164027501236004.

Pinheiro, M. C. and da Silva, F. M. (2010). Comunicação visual e design inclusivo, cor, legibilidade e visão envelhecida. *Design Ergonômico – Estudos e Aplicações*, 17033:62.

Pur, D. R., Lee-Wing, N., and Bona, M. D. (2023). The use of augmented reality and virtual reality for visual field expansion and visual acuity improvement in low vision rehabilitation: a systematic review. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 261(6):1743–1755. DOI: https://doi.org/10.1007/s00417-022-05972-4.

Qiu, S., Hu, J., Han, T., and Rauterberg, M. (2024). Can blindfolded users replace blind ones in product testing? an empirical study. *Behaviour & Information Technology*, 43(8):1664–1682. DOI: https://doi.org/10.1080/0144929X.2023.2226768.

Rahimi, F., Sadeghi-Niaraki, A., and Choi, S.-M. (2025). Generative ai meets virtual reality: A comprehensive survey on applications, challenges, and future direction. *IEEE Access*. DOI: https://doi.org/10.1109/ACCESS.2025.3574779.

Raybourn, E. M., Stubblefield, W. A., Trumbo, M., Jones, A., Whetzel, J., and Fabian, N. (2019). Information design for xr immersive environments: Challenges and opportunities. In *International Conference on Human-Computer Interaction*, pages 153–164. Springer. DOI: https://doi.org/10.1007/978-3-030-21607-8_12.

Real, S. and Araujo, A. (2019). Navigation systems for the blind and visually impaired: Past work, challenges, and open problems. *Sensors*, 19(15):3404. DOI: https://doi.org/10.3390/s19153404.

Rodrigues, C. S. C., Nazareth, V., Azevedo, R. O., Barbosa, P., and Werner, C. (2025). Unseen: Advancing digital accessibility with binaural audio technology in an immersive gaming prototype. *Journal on Interactive Systems*, 16(1):98–108. DOI: https://doi.org/10.5753/jis.2025.4439.

Rubio-Tamayo, J. L., Gertrudix Barrio, M., and García García, F. (2017). Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal technologies and interaction*, 1(4):21. DOI: https://doi.org/10.3390/mti1040021.

Runeson, P. and Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131–164. DOI: https://doi.org/10.1007/s10664-008-9102-8.

Shen, L., Shen, E., Luo, Y., Yang, X., Hu, X., Zhang, X., Tai, Z., and Wang, J. (2022). Towards natural language interfaces for data visualization: A survey. *IEEE transactions on visualization and computer graphics*, 29(6):3121–3144. DOI: https://doi.org/10.1109/TVCG.2022.3148007.

Silva, G. M. S., de C. Andrade, R. M., and de Gois R. Darin, T. (2019). Design and evaluation of mobile applications for people with visual impairments: A compilation of usable accessibility guidelines. In *Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems*, pages 1–10. DOI: https://doi.org/10.1145/3357155.3358450.

Silveira, C. S., dos Santos, A. N., dos Santos Oliveira, C., and de Souza Arnaut, F. F. (2024). Development of accessible virtual reality technology for user inclusion. *JOURNAL OF BIOENGINEERING, TECHNOLOGIES AND HEALTH*, 7(Suppl2):15–22. DOI: https://doi.org/10.34178/jbth.v7iSuppl2.447.

Story, M. F. (2001). Principles of universal design. *Universal design handbook*, 2.

Wang, Z. and Ben-Arie, J. (1996). Conveying visual information with spatial auditory patterns. *IEEE transactions on speech and audio processing*, 4(6):446–455. DOI: https://doi.org/10.1109/89.544529.

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*.

Whicher, D., Philbin, S., and Aronson, N. (2018). An overview of the impact of rare disease characteristics on research methodology. *Orphanet journal of rare diseases*,

13(1):14. DOI: https://doi.org/10.1186/s13023-017-0755-5.

Yuan, B., Folmer, E., and Harris Jr, F. C. (2011). Game accessibility: a survey. *Universal Access in the information Society*, 10(1):81–100. DOI: https://doi.org/10.1007/s10209-010-0189-5.

Zhao, Y., Bennett, C. L., Benko, H., Cutrell, E., Holz, C., Morris, M. R., and Sinclair, M. (2018). Enabling people with visual impairments to navigate virtual reality with a haptic and auditory cane simulation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14. DOI: https://doi.org/10.1145/3173574.317369.

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. DOI: https://doi.org/10.1145/3290605.3300341.