# EnVisionVR: A Scene Interpretation Tool for Visual Accessibility in Virtual Reality

Junlong Chen, Rosella P. Galindo Esparza, Vanja Garaj, Per Ola Kristensson, John Dudley

*Abstract*—**Effective visual accessibility in Virtual Reality (VR) is crucial for Blind and Low Vision (BLV) users. However, designing visual accessibility systems is challenging due to the complexity of 3D VR environments and the need for techniques that can be easily retrofitted into existing applications. While prior work has studied how to enhance or translate visual information, the advancement of Vision Language Models (VLMs) provides an exciting opportunity to advance the scene interpretation capability of current systems. This paper presents EnVisionVR, an accessibility tool for VR scene interpretation. Through a formative study of usability barriers, we confirmed the lack of visual accessibility features as a key barrier for BLV users of VR content and applications. In response, we used our findings from the formative study to inform the design and development of EnVisionVR, a novel visual accessibility system leveraging a VLM, voice input and multimodal feedback for scene interpretation and virtual object interaction in VR. An evaluation with 12 BLV users demonstrated that EnVisionVR significantly improved their ability to locate virtual objects, effectively supporting scene understanding and object interaction.**

*Index Terms*—**Virtual Reality (VR), Vision Language Models, Visual Accessibility, Blind and Low Vision Users.**

## I. Introduction

Virtual Reality (VR) is a primarily visual medium. The centrality of visual perception in the VR experience presents a major challenge when making the technology accessible to Blind and Low Vision (BLV) users. While screen readers and audio descriptions have played a crucial role in enabling BLV users to access information from two-dimensional (2D) screens, this accessibility issue persists for three-dimensional (3D) spatial content. In contrast with how screen readers and audio descriptions work on conventional 2D user interfaces, the current form of VR applications challenges the systematic organisation and delivery of 3D spatial information in an intuitive and efficient format.

In an effort to address the exclusion of BLV users from VR experiences, prior work has studied visual accessibility design in virtual [47, 46] and augmented reality [19]. Wang et al. [40] also studied VR accessibility practices among VR developers. These efforts seeking to improve VR visual accessibility have adopted different strategies, such as enhancing visual information through view magnification, brightness/contrast adjustment, object contour highlighting [47]; or converting visual information to other forms like audio descriptions of virtual objects [47] or vibrotactile feedback [46]. With the advent of Vision Language Models (VLMs), new opportunities are emerging to generate vivid and detailed scene descriptions based on the user's field of view. Such capability can be embedded in output modalities such as speech, audio, and haptic cues to facilitate the user's understanding of 3D scenes.

This paper presents EnVisionVR, an integrated set of VR scene interpretation and virtual object localization tools that assist BLV users in navigating VR. The development was guided by a formative usability study with eight BLV participants, which provided empirical information on the support required by BLV users and, particularly, the types of accessibility features that support scene understanding and interaction. EnVisionVR was then implemented as a proof-of-concept system to improve visual accessibility in VR by providing (a) high-level natural language scene interpretation powered by a VLM, and (b) detailed low-level object description and localization tools based on speech, audio, and haptic cues. The system was evaluated in a user study with 12 BLV participants, who were asked to complete three tasks related to scene understanding, object localization, and object interaction with and without EnVisionVR in a VR scene. Participants achieved a significantly higher success rate when locating virtual objects with EnVisionVR.

This research makes three main contributions. First, the formative study adds to the existing literature on accessibility barriers for BLV users by emphasizing the lack of functions for scene description and interaction support as a key concern. Second, to the best of our knowledge, EnVisionVR is the first proof-of-concept system to incorporate detailed VLM-based scene descriptions for real-time visual accessibility in VR, through spatial audio, voice instructions, and speech-based function activation methods. Third, we offer a set of design implications derived from the system's development process and evaluation to inform visual accessibility design in VR more extensively.

## II. Related Work

### A. Visual Accessibility Design in VR

In a study conducted by Naikar et al. [34], 39 out of 106 inspected free VR experiences (36.8%) lacked accessibility features. Furthermore, users may encounter multiple accessibility barriers in the same context [11]. Anderton et al. [2] categorized accessibility features in 330 VR applications, while other works focused more specifically on visual accessibility in VR to provide a more inclusive experience [10, 15]. Mostly, this has been approached through augmenting visual

Email: jc2375@cam.ac.uk
Email: rosellapaulina.galindoesparza@brunel.ac.uk
Email: vanja.garaj@brunel.ac.uk
Email: pok21@cam.ac.uk
Email: jjd50@cam.ac.uk

information [47, 31, 38] or translating it into audio or haptic feedback [46, 25, 47, 22, 28].

Outstanding work in the area of augmenting visual information includes the development of tools for magnification, contrast adjustment, color correction, text and display size adjustment, among others. Gear VRF Accessibility [38], for instance, provided a framework for developers to adapt zoom, invert colors, and add captions in a VR environment. VRiAssist [31] supported the user by offering visual assistance based on eye tracking, providing tools like magnification, distortion, color and brightness correction. SeeingVR [47] involved a larger set of visual augmentation tools that proved effective for task completion in VR (such as menu navigation, visual search, and target shooting). Consistent with these approaches, Ciccone et al. [9] recommended implementing contrast adjustment controls, color correction controls, and font and display size adjustments to increase information visibility when designing for visual accessibility.

Research focused on converting visual information into other forms has also resulted in a variety of systems supporting visual accessibility in VR. For instance, both Canetroller [46] and VIVR [25] simulated the use of a white cane in the virtual world. This included providing 3D spatial audio feedback, physical resistance, and vibrotactile feedback to simulate cane–virtual object interaction. The aforementioned SeeingVR [47] also included text-to-speech and object recognition from visual information to speech. In a more specialized context, Dang et al. [12] outlined a multimodal-multisensor VR system with spatial audio, audio descriptions, audio feedback, and vibrotactile feedback to enhance the experience of BLV participants in immersive musical performances. Lança et al. [28] studied different techniques to communicate the position of buttons on a grid and found speech to be more intuitive over sonification in sharing the 2D grid position. Finally, VRBubble [22] enhanced BLV users' peripheral awareness to facilitate social VR accessibility through audio alternatives such as earcons, verbal notifications, and real-world sound.

Among both approaches, augmenting visual information cannot support users who are blind or with very limited visual perception. Thus, the work in this paper focuses on integrating the relatively underexplored methods of converting visual information into speech, audio cues, and haptics. We investigate how VLMs could be incorporated to provide vivid scene descriptions. By combining these multiple modalities, we aim to provide users with a high-level understanding of their surroundings, as well as a detailed understanding of object-level information to support interaction.

### B. Screen Readers and Web Accessibility

Screen readers are a well-established accessibility tool for BLV users; their design concepts can provide valuable insights for the design of visual accessibility in immersive environments. NVDA, JAWS, and VoiceOver are three of the most commonly used screen readers for desktops and laptops [41]. While these different screen readers have distinct characteristics, they share key design principles which underpin their effectiveness. First, popular screen readers prioritize keyboard navigation. Keyboard navigation allows users to navigate digital content without the need for a mouse, which is critical for people with vision impairment [24]. Second, screen readers focus on the semantic structure to facilitate smooth navigation and ensure information accuracy. On this topic, a series of works [48, 14, 42] have specifically focused on how to improve the usability of screen readers by correctly and efficiently conveying semantic details. Third, screen readers also provide alternative text for images, which is a crucial step to help convey non-textual content [43, 33]. Fourth, screen readers use headings and landmarks to assist website navigation and hierarchy [37]. Finally, screen readers also assist user input, such as filling in and submitting forms and documents online, an important part of web interaction [6].

EnVisionVR takes inspiration from and expands on the design principles and concepts of screen readers and audio descriptions. Based on the above, we arrive at an interactive design that uses speech commands as a parallel to keyboard navigation, while constructing high-level scene information and detailed object-level information for BLV users as a parallel to the semantic structure processed by screen readers. Furthermore, VLMs provide a highly efficient way to produce audio scene descriptions, a parallel to explicit alternative text.

### C. Powering Visual Accessibility with Artificial Intelligence

The emergence of powerful VLMs has enabled the automated generation of high-quality descriptions of visual information. Current VLMs [35, 8, 30, 5, 45] are capable of jointly processing images and text data for image captioning, visual question answering, and medical image analysis. These models are now being deployed in a range of use cases to power visual accessibility features. For example, De La Torre et al. [13] demonstrated potential applications of their Large Language Model (LLM)-based tool for 3D scene editing in visual accessibility. Jiang et al. [23] highlighted the potential of advanced AI models to enhance the quantity and quality of audio descriptions.

For physical-world scenarios, Microsoft developed SeeingAI [32] to narrate the physical world for BLV users. Similarly, Be My Eyes launched Be My AI [16], an AI assistant powered by GPT-4, which provides detailed descriptions of photos taken and uploaded by BLV users, and a braille display for deaf-blind users. Vision-language models like WorldScribe [7] and multimodal large language models like VIAssist [44] generate live visual descriptions of the real world for BLV users. Kuribayashi et al. proposed WanderGuide [27], a robotic guide which assists recreational indoor exploration for blind users by providing different levels of description detail and verbal interaction. Specific use cases for scene description in real-life scenarios have been identified through a diary study [17], which highlights the effectiveness of generative models for visual accessibility design.

The increasing attention to applying VLMs to interactions in 3D content and accessibility design illustrates the strong capability of such models, but there has been limited work studying how these models could be applied in accessibility design for VR immersive environments. Our work addresses this gap by leveraging the capabilities of VLMs for VR scene interpretation. Existing physical-world vision-language pipelines [32, 16, 7, 44, 27] highlight key implications including the need for context-aware descriptions, spatialized cues, and user-centered navigation support, which inform the design of VLM-assisted visual accessibility systems in immersive environments.

## III. Formative Study

In the formative usability study, which involved eight BLV participants, we sought to understand the accessibility barriers encountered in consumer-based VR and AR technology. Through this process, we studied the adaptations implemented when facing such barriers, namely the way people with specific access needs modified their behaviour or received assistance from another non-disabled person to fully or partially overcome these issues.

### A. Method

Our study protocol evaluated the usability of representative consumer-level, single-user VR experiences to identify the types of barriers encountered. The majority of the experiences represented content currently available in the market, while others were included to cover the full range of usability demands in VR, such as vision, hearing, touch and physical movement, and interaction modes, such as controller-based and hand-tracking. The study tasks and experiences became progressively more complex as the study progressed. See the Online Appendix for the complete set of experiences, tasks and sub-tasks.

Meta Quest 2 was used. Tasks and experiences were designed following the typical user journey; beginning with wearing and fitting the VR hardware (*VRH: Headset*, *VRC: Controllers*), followed by navigating the Meta Quest universal menu to configure existing accessibility features (*VR1: Menu*), and completing each of the selected experiences: *VR2: "As it is" 360° video*[1] (immersive video documentary), *VR3: Job Simulator*[2] (videogame simulating a cooking scenario, using virtual hands to manipulate objects while following cooking instructions), *VR4: Moss*[3] (storyline-based videogame where the user becomes a secondary character that interacts with objects and controls other characters), and *VR5: Elixir*[4] (hand-tracking-based videogame where the user manipulates virtual objects with their real hands). Sub-tasks were basic commands revolving around specific steps required to progress through each experience and explore available features and interactables.

In total, participants completed 34 sub-tasks spread across two VR hardware tasks and five VR experiences (e.g., *'Adjust the focal distance of the headset'*, *'Spot different visitors in the scene, from those close by to those at a distance'*). Participants were asked to perform each sub-task while thinking aloud. A researcher scored task success on a 0–3 scale (0 = unable to start or finish the task, even with adaptations, 1 = able to start but unable to finish the task, even with adaptations, 2 = successful completion of the task with adaptations, and 3 = successful completion of the task without adaptations). The concept of *adaptation* arose after a pilot study that showed most sub-tasks were not achievable for multiple people with access needs. Thus, we resolved to study adaptations as either *self-initiated* unconventional behaviour (e.g., holding a VR controller with two hands for pointing accuracy) or *assistance* where the researcher supported the participant in achieving their goals by mimicking plausible but unavailable accessibility functionality (e.g., imitating a non-existent screen reader feature).

The study was approved by the Ethics Committee of the College of Engineering, Design and Physical Sciences, Brunel University of London. The study session lasted approximately 120 minutes per participant. As shown in Figure 1, the sessions were facilitated by a researcher experienced in providing BLV accessibility support; they were in charge of observing, scoring sub-tasks, and assisting the participants. A technician was in charge of onboarding and looking after the technical elements.

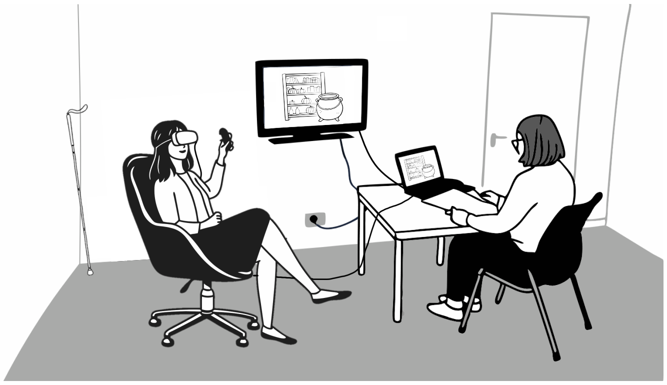


Fig. 1. Setup of the formative and evaluative study.

Eight participants (2 female, 6 male) who self-reported as blind or with low vision were recruited through an inclusive research user panel (managed by Open Inclusion [21]). All participants provided informed consent. Their ages ranged from 27 to 68 (*M* = 43.63 years, *SD* = 13.96) and their previous experience with VR technology ranged from novice (1) to competent (3). For these participants, sight was classed as the access need that impacted their lives most extensively. These details are summarized in Table I. To distinguish from participants in the study reported in Section V, participants in the formative study are labelled PF1 to PF8.

### B. Results

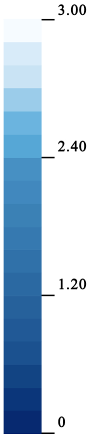
*1) Task Success:* This score indicates the level of success in completing a sub-task. Each participant was presented with 34 sub-tasks in total (*VRH: Headset* = 4, *VRC: Controllers* = 6, *VR1: Menu* = 6, *VR2: 360° video* = 7, *VR3: Job Simulator* = 4, *VR4: Moss* = 5, *VR5: Elixir* = 2). 271 individual scores were produced across the eight participants over the seven VR tasks/experiences; one sub-task was not performed

[1] Produced by 360 Labs
[2] Produced by Owlchemy Labs
[3] Produced by Polyarc
[4] Produced by Magnopus

TABLE I

PARTICIPANT DEMOGRAPHICS AND TASK SUCCESS SCORES IN THE FORMATIVE STUDY. VR EXPERTISE INDICATES SELF-REPORTED EXPERIENCE WITH VR ON A SCALE FROM 1 (NOVICE) TO 5 (EXPERT). MEAN (M) AND STANDARD DEVIATION (S) ARE REPORTED FOR EACH TASK OR EXPERIENCE.

| | | | | | | Task Success Scores | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Vision | VR Expertise | | VRH | VRC | VR1 | VR2 | VR3 | VR4 | VR5 |
| Participant | Age | Gender | Description | (1-5) | | Headset | Controllers | Menu | 360° Video | Job Sim | Moss | Elixir |
| PF1 | 56 | F | Blind | 2 | *M* | 2.75 | 2.83 | 0.50 | 1.00 | 0.75 | 1.20 | 1.00 |
| | | | | | *S* | 0.50 | 0.41 | 0.84 | 1.29 | 0.50 | 0.45 | 0.00 |
| PF2 | 38 | F | Lacks most central vision and right peripheral vision. Nystagmus | 3 | *M* | 3.00 | 3.00 | 1.50 | 2.57 | 2.50 | 2.40 | 2.50 |
| | | | | | *S* | 0.00 | 0.00 | 1.22 | 1.13 | 1.00 | 0.55 | 0.71 |
| PF3 | 41 | M | Blind | 2 | *M* | 3.00 | 3.00 | 0.00 | 0.43 | 1.75 | 2.00 | 1.00 |
| | | | | | *S* | 0.00 | 0.00 | 0.00 | 1.13 | 1.26 | 0.00 | 0.00 |
| PF4 | 56 | M | Blind in left eye, only central vision in right eye | 1 | *M* | 3.00 | 3.00 | 2.00 | 2.43 | 2.50 | 2.20 | 2.00 |
| | | | | | *S* | 0.00 | 0.00 | 0.00 | 0.53 | 0.58 | 0.45 | 0.00 |
| PF5 | 68 | M | Double vision due to multiple sclerosis | 1 | *M* | 2.00 | 3.00 | 3.00 | 3.00 | 1.75 | 2.20 | 2.00 |
| | | | | | *S* | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.45 | 0.00 |
| PF6 | 27 | M | Irlen syndrome | 1 | *M* | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| | | | | | *S* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| PF7 | 32 | M | Blind | 2 | *M* | 3.00 | 3.00 | 0.00 | 0.43 | 2.00 | 1.20 | 1.50 |
| | | | | | *S* | 0.00 | 0.00 | 0.00 | 1.13 | 0.00 | 0.45 | 0.71 |
| PF8 | 39 | M | Aniridia with sensitivity to light and glare. Nystagmus | 1 | *M* | 3.00 | 3.00 | 3.00 | 2.14 | 2.75 | 2.40 | 2.00 |
| | | | | | *S* | 0.00 | 0.00 | 0.00 | 0.90 | 0.50 | 0.55 | 0.00 |
| | | | | TOTAL (8) | *M* | 2.84 | 2.98 | 1.63 | 1.88 | 2.13 | 2.08 | 1.88 |
| | | | | | *S* | 0.18 | 0.14 | 0.49 | 0.52 | 0.43 | 0.23 | 0.33 |
| | | | | | | VRH | VRC | VR1 | VR2 | VR3 | VR4 | VR5 |



due to participant request (PF7, *VRH-4*). Mean scores are summarized in Table I. The *Menu* (*VR1*) presented the lowest overall mean score while the highest was observed in *Job Simulator* (*VR3*). While low-vision participants were generally able to fully or partially complete the sub-tasks in the *Menu* (*VR1*) and *360° video* (*VR2*), blind participants were largely unable to initiate these tasks. This gap narrowed in subsequent experiences (*VR3: Job Simulator*, *VR4: Moss* and *VR5: Elixir*), although blind participants continued to score on the lower end of the scale overall.

*2) VR Accessibility Barriers:* Whenever a participant scored 2 or less in Task Success in a sub-task, a usability friction instance was logged and analyzed using thematic analysis. A total of 127 instances of usability friction were encountered. Most frictions were linked to a single barrier type (n = 109), with others involving two (n = 17) or three types (n = 1), totaling 146 barriers across 16 categories (see Online Appendix). The most frequent barrier (32.19%, n = 47) concerned challenging interactions in virtual environments relying solely on visual cues (e.g, low-contrast graphic indicators). This was followed by spatial navigation difficulties due to absent spatial audio, tactile guidance, and visual landmarks, and limited non-visual feedback for menu operation (14.38%, n = 21 each). Of the 127 recorded frictions, 111 were resolved through facilitator assistance, often involving a combination of strategies. The most frequent support included providing ad-hoc audio descriptions (n = 66) and guidance to direct the participant (n = 65). In some cases, facilitators read on-screen text to compensate for missing screen reader functionality (n = 34).

*3) Adaptations for Blind Participants:* Blind participants had difficulty with experiences that only provided single-modality outputs. This issue was particularly notable when information was communicated solely through visual means, but participants also faced challenges interpreting information provided in a single modality using either audio or haptics. Audio descriptions of the play space, interactable objects and pointer location were often necessary. This was more common at the start of an experience and when haptic or audio cues alone failed to convey object types (PF1, PF3, PF7).

Verbal guidance was helpful when friction occurred. On most occasions, the facilitator guided participants on controller use, for example, to explain how controllers were mapped to interactions in a specific scene, or how to manipulate interactables or control characters (PF1, PF3, PF7). In this regard, PF3 highlighted the need for a directional cueing system that could, for instance, guide them to move their controllers closer to the menu. Audio and haptic cues combined were another requirement identified throughout the study. When they were provided conjointly (e.g., *VR3: Job Simulator* used haptics and audio to simulate the opening of a virtual door), PF1 and PF3 could more easily perceive what was going on in the scene. When such signals were poor or did not exist, it became more difficult for participants to orient themselves (PF1, PF3, PF7).

*4) Adaptations for Low-Vision Participants:* Low-vision participants, such as PF4, completed more sub-tasks than blind participants but required longer periods to familiarize themselves with the virtual environments. Audio descriptions and verbal guidance were important to clarify what participants were partially seeing in a scene. PF8, for instance, benefited from audio description in the *360° video* (*VR2*). Detailed and repeated instructions were helpful for PF2 in *Elixir* (*VR5*). Existing multimodal feedback was helpful in some instances. For example, multiple signals (i.e. peripheral vision, haptics and sound) helped PF2 manipulate the interactable objects in *Job Simulator* (*VR3*). However, the *Menu's (VR1)* multimodal feedback did not suffice. In this case, PF2 struggled operating it because the haptic feedback did not confirm specific button interaction. Similarly, in *Moss* (*VR4*), several participants struggled due to low color contrast and unclear audio indicators of object interactivity (PF2, PF4, PF8).

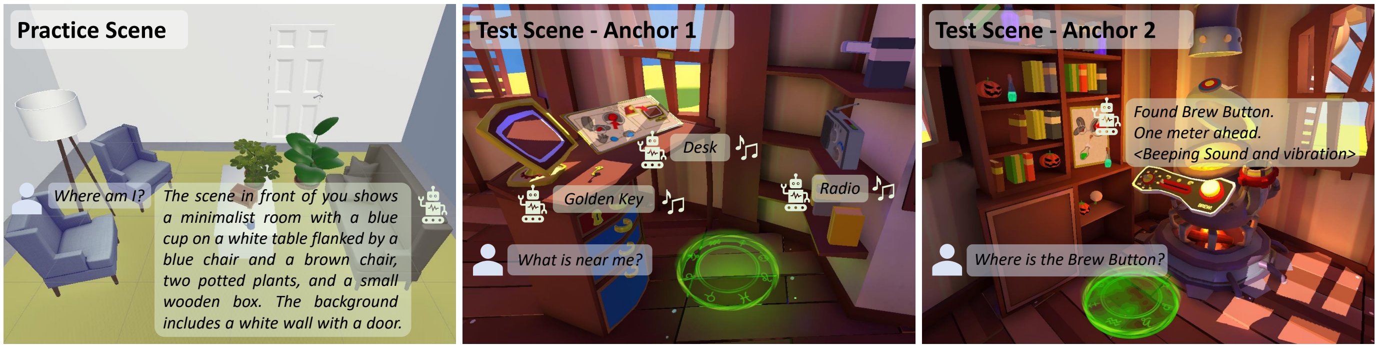

Practice Scene
Where am I?
The scene in front of you shows a minimalist room with a blue cup on a white table flanked by a blue chair and a brown chair, two potted plants, and a small wooden box. The background includes a white wall with a door.
Test Scene - Anchor 1
Desk
Golden Key
Radio
What is near me?
Test Scene - Anchor 2
Found Brew Button.
One meter ahead.
<Beeping Sound and vibration>
Where is the Brew Button?

Fig. 2. Examples of functions supported by EnVisionVR to enhance the accessibility of VR experiences for BLV users. Left: The user can ask "Where am I?" and the EnVisionVR system reads out a detailed description of the user's current field of view. Middle: The user can ask "What is near me?" and the system reads out the names of the three main objects near the user with a spatial tone to indicate the object's location. Right: The user can ask "Where is the Brew Button?" and the system uses a beeping sound and directional instructions to communicate the distance to the Brew Button. When the user reaches the Brew Button, the controller vibrates to inform the user.

*5) Accessibility Feature Priorities:* Through the think-aloud process, participants consistently highlighted the need for multimodal feedback to overcome the limitations of VR environments only providing visual cues. This was supported by the facilitator's observations. Key recommendations included integrating clear audio and haptic signals to indicate interactable objects and menus (PF1, PF3); for example, distinct audio tones should differentiate interface elements, while haptic feedback can confirm successful interactions. Continuous environment and object audio descriptions were recommended by the facilitator, particularly during initial orientation and when other cues proved insufficient. Additionally, integrating screen reader compatibility for all on-screen text, including menus and subtitles, was identified by the facilitator as crucial BLV support. Desired features to avoid spatial orientation and navigation difficulties included a directional cueing system guiding hand controllers to interactive elements (PF3), expanded menu targeting ranges (PF3, PF4, PF6, PF8), and the use of binaural or 3D audio to convey spatial depth (PF1, PF3, PF4). Additional environmental audio cues were requested to enhance immersion and provide crucial feedback (PF1, PF3).

### C. Summary

Results from the formative study revealed that BLV users face various accessibility barriers in VR. Notably, participants were unable to complete tasks when interactions relied solely on visual cues. Audio and haptic signals which were insufficient to convey spatial layout or object interactivity also prevented participantsThe lack of integrated screen reader functionality and audio descriptions were identified as major barriers preventing them from completing sub-tasks without assistance. While low-vision participants required longer periods to familiarize themselves with the environments, or concrete audio descriptions to clarify the scenes, blind participants were unable to carry out specific sub-tasks in *Moss* (*VR4*) and *Elixir* (*VR5*) because there was no appropriate multimodal feedback to, for instance, understand the location of objects. In contrast, *Job Simulator* (*VR3*) offered helpful audio cues when reaching interactable objects.

The provision of accessibility features was irregular across the experiences, consistent with the findings of Naikar et al. [34]. BLV participants continually struggled to complete sub-tasks related to visual capability demands, revealing that where visual accessibility features existed (e.g., colour contrast adjustment, audio levels) these were insufficient. It was also observed that isolated screen reader or audio description implementations are unlikely to accommodate BLV users' requirements. In contrast, a more complex system dedicated to providing high-level descriptions of the virtual environments coupled with detailed descriptions of interactable objects could serve as an effective guide both for blind and low-vision users facing difficulty with navigating VR experiences.

## IV. Design of EnVisionVR

The formative study suggests that BLV users require a scene interpretation functionality which builds upon the principles of screen readers and audio descriptions to incorporate spatial elements to assist users to navigate and interact within the 3D space. While prior work, such as SeeingVR [47], has focused on features that support visual perception (e.g. zooming, contour highlighting), we pursue an alternative strategy and seek to replicate and evaluate the familiar experience of using a screen reader and listening to audio descriptions to provide visual accessibility for 3D content.

In response to findings in the formative study, we developed EnVisionVR, a generalized visual accessibility framework for VR applications. Table II provides a summary of how findings from the formative study informed the design of EnVisionVR. This framework consists of: (i) the Scene Description Function; (ii) the Main Objects Indication Function; and (iii) the Object Localization Function. To minimize the need for remapping controller buttons, these functions can be activated by three simple speech commands, namely "Where am I?", "What is near me?", and "Where is the <object name>?". For the implementation evaluated in this paper, the user must first press Button A on the right handheld controller to issue a voice command but this could in theory be changed to a 'wake' word or remapped to any other button. The use of these three functions is illustrated in Figure 2 and their

TABLE II
EnVisionVR design goals based on formative study findings.

| Formative Finding | EnVisionVR Design Goals |
|---|---|
| Multimodal feedback overcomes visual-only limitations | EnVisionVR should convey information via various modalities such as speech, audio, and haptic vibrations. |
| Continuous descriptions enable scene understanding | The system should support real-time scene descriptions along with follow-up descriptions for more precise object queries. |
| Binaural and 3D audio convey spatial depth | Spatialized audio should guide users toward objects with speech and/or audio cues. |
| Environmental audio enhance immersion and provide feedback | EnVisionVR should enhance immersion and address user needs for rich auditory feedback by providing features such as scene-level descriptions and object-level audio cues. |

implementation is described in more detail in the remainder of this section.

## A. *Scene Description – Where am I?*

Issuing the **"Where am I?"** voice command triggers the **Scene Description Function**, which describes the user's field of view in a few sentences. In line with Iachini et al. [20], egocentric descriptions of the scene were provided instead of allocentric descriptions to improve the cognition of BLV users in large room-scale spaces. An overview of the implementation of the Scene Description Function before and during runtime is provided in Figure 3. Details of each step in the implementation are provided in the Online Appendix.

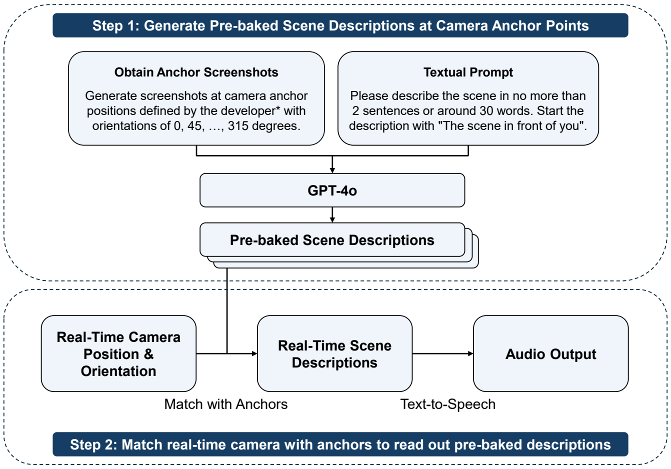

Step 1: Generate Pre-baked Scene Descriptions at Camera Anchor Points
Obtain Anchor Screenshots
Generate screenshots at camera anchor positions defined by the developer* with orientations of 0, 45, ..., 315 degrees.
Textual Prompt
Please describe the scene in no more than 2 sentences or around 30 words. Start the description with "The scene in front of you".
GPT-4o
Pre-baked Scene Descriptions
Real-Time Camera Position & Orientation
Real-Time Scene Descriptions
Audio Output
Match with Anchors
Text-to-Speech
Step 2: Match real-time camera with anchors to read out pre-baked descriptions

Fig. 3. Overview of the Scene Description Function. Scene description is provided in two steps. In Step 1, camera anchor positions are determined by the developer or automatically by the system. Screenshots of the field of view of these anchor points with orientations of 0, 45, ..., 315 degrees along the horizontal plane together with a textual prompt are fed into GPT-4o to generate pre-baked scene descriptions. In Step 2 during runtime, we match the current camera position and orientation with the closest-matching anchor position and orientation to read out the pre-baked descriptions via the Microsoft text-to-speech (TTS) service.

Before runtime, camera anchor points[5] are determined manually by the developer. As shown in Figure 4, upon specifying the camera anchor points, a script is executed to automatically capture eight screenshots of the user field of view at each anchor point with orientations of 0, 45, ..., 315 degrees. These screenshots are then sent to a vision language model (VLM) together with a textual prompt, and a short scene description is obtained for each camera anchor position and preset orientation. For example, if four camera anchor positions are determined, $4 \times 8 = 32$ scene descriptions are generated. We used GPT-4o as the VLM and used the textual prompt "Please describe the scene in no more than 2 sentences or around 30 words. Start the description with 'The scene in front of you"' to generate the scene descriptions. The scene descriptions are stored locally in a CSV file. As the scene descriptions are generated before users execute the application, we refer to these descriptions as *'pre-baked'*[6].

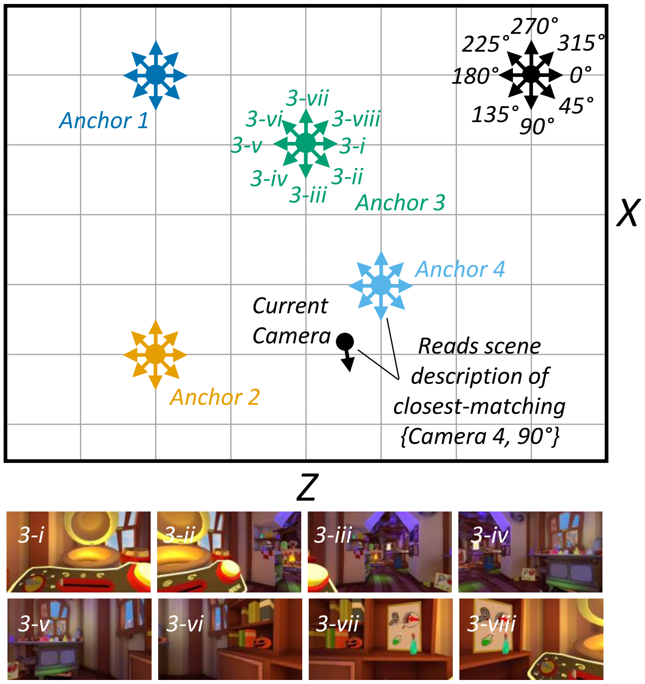


Fig. 4. Top-down view of camera anchor positions in a VR escape room and the user field of view in eight directions for each anchor point. At each field of view, a screenshot is taken to generate the pre-baked scene description. Example field-of-view screenshots taken at Anchor 3 are provided in the bottom images.

During runtime, the Scene Description Function processes the current user position and orientation to find the anchor point with the closest matching position and orientation and reads out its 'pre-baked' scene description using the Microsoft Azure text-to-speech (TTS) service. As the scene descriptions have been generated by the VLM before runtime, this mapping

[5] We define *anchor points* as a list of $(x, y, z)$ coordinates which define the position, but not the orientation, for the user camera to be placed in the scene, such that the user camera placed at all anchor points, with eight different orientations each, capture user field of views with all of the important objects in the scene.

[6] Currently, scene descriptions are generated prior to users executing the application (i.e. pre-baked). If the VR scene is modified during runtime the scene descriptions will not be accurate.

process allows scene descriptions to be read out to the user with very low latency, usually within tens of milliseconds ($M = 19.8, SD = 19.5$) based on data from our user study.

### *B. Main Objects Indication – What is near me?*

Issuing the **"What is near me?"** voice command triggers the **Main Objects Indication Function**, which announces the names of three key objects near the user (not necessarily in the field of view), each followed by a short spatial tone to indicate the object's location relative to the user headset. Exactly which objects are read out is determined by a 'runtime importance value'. This value is proportional to a preset importance value, and inversely related to the distance between the object and the user camera. Here, the preset importance value for all objects can be determined automatically by ENVISIONVR based on the presence of rendering components (such as Mesh Renderer components attached to the game object in Unity), or specified manually by the developer. If an object has been previously announced, its runtime importance value is reduced to allow other objects to be announced in subsequent activations of the function. Further details are provided in the Online Appendix.

### *C. Object Localization – Where is the <object name>?*

Issuing the **"Where is the <object name>?"** voice command triggers the **Object Localization Function**, which starts a beeping sound with the beeping frequency inversely proportional to the distance between the right controller and the object. Directional and distance information (e.g., '1 meter ahead') is also provided at regular time intervals to guide the user. When the controller is close enough to the object to interact with it, the controller vibrates. If the object is interactable and can be held, the system will also announce "holding <object name>" when it is picked up. More implementation details are provided in the Online Appendix.

## V. EVALUATION STUDY

To evaluate the potential benefits of ENVISIONVR, we conducted a user study with 12 BLV participants. Participants completed three types of prescribed tasks in VR both with ENVISIONVR and without. The without condition represented the default experience available to BLV users with no dedicated visual accessibility features. The study was approved by the research ethics committee in the Department of Engineering at the University of Cambridge.

### *A. Method*

A within-subjects design was adopted to evaluate the performance of ENVISIONVR (abbreviated in Section VI as EVR) and the no accessibility features condition (abbreviated in Section VI as NVR). The order of conditions was counterbalanced. For each condition, participants were first familiarized with the available functions in a practice scene (see Figure 2 left image). Participants were encouraged to use all available functions and the experimenter gave examples of the types of tasks they would be asked to complete. After completing familiarization, participants were transported to the test scene.

The test scene, a VR Escape Room [39], featured familiar objects (e.g., table, key, shelf) as well as several fantasy objects (e.g., cauldron, potion bottles) arranged within a wooden cabin. This scene was chosen for several key reasons. First, it is a tutorial scene made freely available by Unity and so provides an example of the type of existing VR experience that may require retrofitting of accessibility features. Second, it contains rich objects and scene elements. Third, it supports different interactions with virtual objects (such as grabbing objects and pressing buttons). We define two study anchor points in this scene (see Figure 2 middle and right image) and the combination of condition and anchor point is balanced across participants. At the given study anchor point, participants then completed the tasks described in the following subsection.

### *B. Tasks and Measures*

The degree to which ENVISIONVR supports BLV users in perceiving and interacting with the virtual environment is evaluated across the three tasks summarized below. A complete list of questions and tasks for the two anchor points in the test scene is provided in the Online Appendix.

1) **Scene Understanding Task:** Participants are asked to rate a statement to evaluate their understanding of the scene from 1 (very unlikely to be true) to 5 (very likely to be true). For example, at Anchor 1 (see Figure 2 middle image), participants are asked to judge whether the statement "This is a scene of a classroom with a desk and a chair" is likely to be true or not.
2) **Object Localization Task:** Participants are asked to turn to face a specified object in the scene, such that the object is in the field of view but not necessarily directly in front of the user. For example, they are asked to turn to face the radio at Anchor 1. The task completion status was recorded as a yes/no binary value.
3) **Object Interaction Task:** Finally, participants are asked to interact with an object in the scene. For example, they are asked to push the "Brew Button" at Anchor 2 (see Figure 2 right image). Again, we record their task completion status as a binary value.

Participants also rated the difficulty they encountered in each task from a scale of 1 to 5. The performance measures for all three tasks above, together with the perceived difficulty, number of function activations, and qualitative comments from post-study interviews form the measures captured in the evaluative study. These are subsequently analyzed in Section VI.

### *C. Participants*

We recruited a new participant sample with the assistance of Open Inclusion [21]. All participants provided informed consent. The sample consisted of 12 participants, of which three reported being blind and nine reported having low vision. All three blind participants reported regular use of screen readers or other forms of assistive technology. Among the nine participants who reported having low vision, five participants reported regular use of assistive technology, yielding a total of eight participants who regularly use assistive technology.

TABLE III
Participant demographics for the evaluative study with BLV users.

| Participant | Age | Gender | Education | VR Experience | Vision | From Birth | Vision Description | Assistive Technology | Regular Use |
|---|---|---|---|---|---|---|---|---|---|
| P1 | 58 | Female | Masters | Inexperienced | Blind | No | Sighted in the past but have no usable vision today. | Voiceover and Jaws as screen readers; Other tech with audio assistance at home. | Yes |
| P2 | 21 | Female | A levels | Inexperienced | Low Vision | Yes | Born with cataracts and glaucoma, able to see a decent amount with glasses. | Uses phone and screen magnifiers to zoom in. | No |
| P3 | 73 | Male | GCSE | Highly Inexperienced | Blind | No | Lost sight gradually, totally blind for the past 2 years. | Uses screen reading software: JAWS, NVDA, Voiceover. | Yes |
| P4 | 50 | Female | Higher National Diploma | Inexperienced | Low Vision | No | Stargardt's which affects central vision. | Uses Voice Over and Zoom Text. | Yes |
| P5 | 79 | Male | GCSE | Highly Inexperienced* | Blind | No | Lost sight during a degree course. | Siri, Alexa, Be My Eyes, JAWS, Voiceover, and other screen readers. | Yes |
| P6 | 36 | Male | College | Neither inexperienced nor experienced | Low Vision | N/A | Can see 1 meter ahead, central vision in one eye only. | Phone has voice over - Apple iPhone. Windows PC, Samsung tablet. Talking TV. Uses Seeing AI - to read bus numbers. | Yes |
| P7 | 58 | Male | Postgraduate degree | Inexperienced | Low Vision | N/A | No sight in left eye, limited central vision (3/60) in right eye. Has ADHD. | Screen magnification user on the computer | No |
| P8 | 36 | Female | AS Level | Highly Inexperienced | Low Vision | N/A | Has light perception and no residual vision. Difficulties in reading if the text is not in the right format. | Uses screen reader on a daily basis. NVDA on laptop. Talkback on Android. Previously iPhone. | Yes |
| P9 | 41 | Male | Bachelor's degree | Highly Experienced | Low Vision | N/A | Little vision in right eye, can see light and dark and the shape of things. | Text enlarger on mobile and computer, and Dragon Naturally Speaking for speech input and feedback | No |
| P10 | 45 | Male | Bachelor's degree | Highly Inexperienced | Low Vision | N/A | Zero sight in left eye, right eye is a prosthetic, 6/36 vision with changing field. Only sees shape and colour. | Has used lots of tech. Doesn't use an actual screen reader. Has reading glasses and uses audiobooks a lot. | Yes |
| P11 | 54 | Female | Bachelor's degree | Highly Inexperienced | Low Vision | N/A | Sight in right eye, no sight in left eye. Born with cataracts. Can read some print with glasses. | Does not use audio on the computer. Does not use specific software. Can enlarge print. | No |
| P12 | 57 | Female | Entry level 2 English | Highly Inexperienced | Low Vision | N/A | No central vision and a tiny bit of peripheral vision. Can see light, dark, and some outlines. | NVDA, Alexa in the house, Android mobile with Synaptec for screen reader. | Yes |

* P5 was new to the concept of VR. He connected VR with soundscapes and gave himself a rating of 'Neither inexperienced nor experienced'. He also said that he had never used technology of this kind later in the testing session, suggesting that an accurate rating could have been 'Highly inexperienced'.

Table III provides a summary of the collected demographic information of all 12 participants. To differentiate from the formative study, participants are labeled as P1 to P12.

### *D. Apparatus*

During the experiment, participants wore a Meta Quest 3 headset and held the right controller. Participants completed all tasks while remaining seated in a swivel chair. The headset was connected to a Windows 11 laptop in wired 'link' mode. Similar to the formative study, a facilitator observed, recorded scores, took notes, and assisted participants, as shown in Figure 1. A second researcher managed technical components. The risks associated with simulator sickness and cognitive burden were mitigated by providing regular experimental breaks and allowing participants to pause or withdraw from the study at any time.

## VI. Results

In this section, we first present the results of 12 participants for each task outlined in Section V-B together with effect sizes and $p$-values. We also report observations of EnVisionVR usage behavior in Section VI-D, and summarize qualitative feedback in post-study interviews in Section VI-E, as recommended for accessibility-focused HCI research [29].

### *A. Scene Understanding Task*

In the Scene Understanding Task, participants responded to the given statement on a scale from 1–"very unlikely" to 5–"very likely". Since at one anchor location, the statement was false, we converted these raw responses such that a higher score indicates a closer match to the correct answer. Figure 5 (left) plots the scene understanding score of all participants in the NVR ($M = 3.67, SD = 1.44$) and EVR ($M = 4.08, SD = .793$) conditions. In Figure 5 we also make a distinction between whether participants regularly use screen readers or other assistive technology. This roughly groups the full participant group into two subsets based on the degree to which they can directly perceive visual content. P3, P4, P5 and P12 who regularly use assistive technology gained a better understanding of the scene with EnVisionVR compared with the condition without any accessibility features. P1, P2, P6, and P10 were able to understand the scene better without EnVisionVR, while P7, P8, P9, and P11 achieved the same level of scene understanding with and without the tool. The decrease in scene understanding performance was due to different reasons such as a lack of attention to long descriptions and failure to capture keywords to support user judgment (P1), or the lack of evidence to convince them to negate the statement which claims the escape room is a classroom (P2, P6). The first-person view descriptions provided

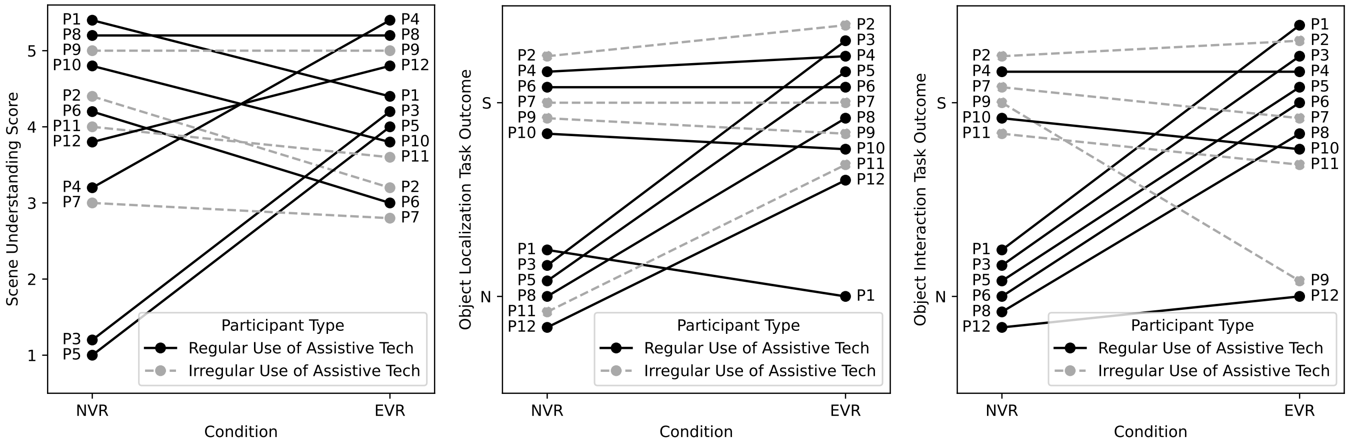


Fig. 5. Performance of all participants in the EVR and NVR condition in the Scene Understanding Task (left), Object Localization Task (middle), and Object Interaction Task (right). Participants with blindness and severe visual impairment who regularly use assistive technology are colored in black, while others are colored in grey. Vertical jittering is applied to visualize all points. Participant IDs are labelled beside each scatter point. S: Successful; N: Not successful.

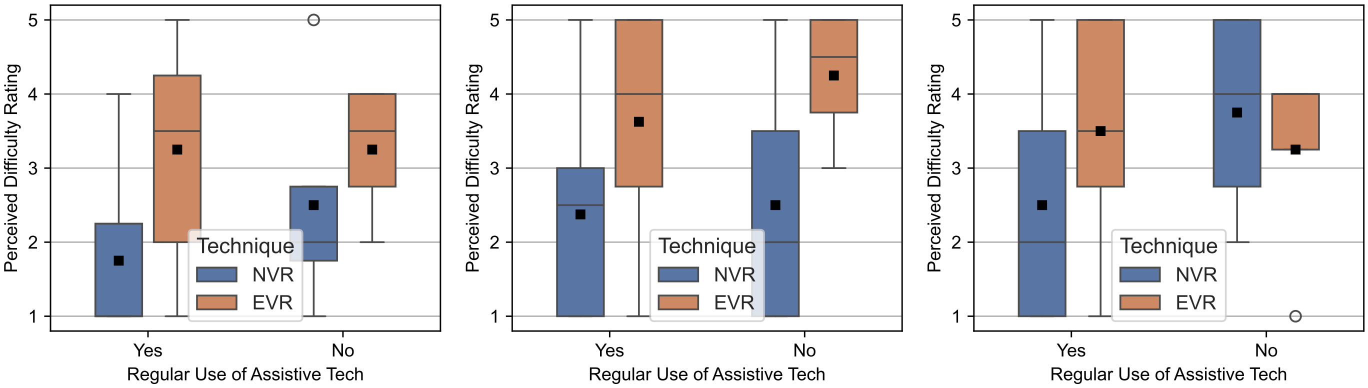


Fig. 6. Distribution of the perceived difficulty (higher score indicates lower perceived difficulty) in the Scene Understanding Task (left), Object Localization Task (middle), and Object Interaction Task (right) for the NVR and EVR conditions for participants who regularly use assistive technology and for those who do not. Black squares indicate the mean value.

only fragments of information about objects around the user, which was insufficient to infer high-level information such as the type and purpose of the scene (P10).

A Wilcoxon Signed-Rank test did not indicate a significant difference in scene understanding scores ($W = 24.0, p = .43, |r| = .33$) between the NVR and EVR conditions. Figure 6 (left) presents boxplots of the perceived difficulty ratings (higher score indicates lower perceived difficulty) of the scene understanding question for the NVR ($M = 2.00, SD = 1.35$) and EVR condition ($M = 3.25, SD = 1.29$). The difficulty ratings are grouped for participants who regularly use assistive technology (NVR: $M = 1.75, SD = 1.16$; EVR: $M = 3.25, SD = 1.49$) and participants who do not (NVR: $M = 2.50, SD = 1.73$; EVR: $M = 3.25, SD = .957$). A Wilcoxon Signed-Rank test did not indicate a significant difference in perceived difficulty ($W = 46.0, p = .06, |r| = .67$) between the NVR and EVR conditions.

## B. Object Localization Task

Figure 5 (middle) summarizes the completion status of the object localization task for all participants. Six participants were able to complete the object localization task to turn to face a specified object in the NVR condition, while the other six participants were not. In the EVR condition, five of the participants who were unable to complete the task in the NVR condition were able to complete the object localization task. Only one participant (P1) was still unable to complete the task in the EVR condition, and none of the participants had a worse performance with EnVisionVR. Most participants (3 out of 4) who do not regularly use assistive technology were able to complete the object localization task with or without EnVisionVR, and EnVisionVR was able to help four out of five participants who do regularly use assistive technology, and who could not locate the virtual object in the NVR condition to complete the task. Overall, object localization task completion results show a $91.7\% - 50\% = 41.7\%$ improvement in task success rate with EVR compared with NVR. As the object localization task has binary performance data, a McNemar's test was adopted. The test indicated a significant difference ($\chi^2 = 5.0, p < .05$, Cohen's $g = .42$) between the NVR and EVR task completion status, suggesting that EnVisionVR significantly improved participants' ability to locate virtual objects.

Figure 6 (middle) presents box plots of the perceived difficulty of the task for the NVR ($M = 2.42, SD = 1.51$) and EVR ($M = 3.83, SD = 1.34$) conditions. A Wilcoxon Signed-Rank test did not indicate a significant difference ($W = 54.0, p = .07, |r| = .64$) between the NVR and EVR condition for all participants.

### C. Object Interaction Task

Figure 5 (right) shows the completion status of the object interaction task. Six participants were able to interact with a virtual object (such as picking up a key or pressing a button) under the NVR condition, while the other six were not. Among those who were unable to interact with virtual objects, five participants were able to complete the task with EnVisionVR, while one participant (P12) was still unable to complete the task. It is worth noting that P12 reported a secondary access need based on her learning disability, and this may have contributed to the difficulty they experienced in completing the task. Among the six participants who were able to complete the interaction task under the NVR condition, five were still able to complete the task with EnVisionVR. However, P9 with little remaining vision was not able to complete the task with EnVisionVR as he felt the main objects indication function provided conflicting information by reporting an object directly behind him, which could not be confirmed easily using vision. Most participants of the subgroup who regularly use assistive technology (6 out of 8) were not able to complete the interaction task in the NVR condition, and EnVisionVR was able to support five out of these six participants to complete the interaction task. Overall, object interaction task completion results show a $83.3\% - 50\% = 33.3\%$ improvement in task success rate with EVR compared with NVR. A McNemar's test did not indicate a significant difference ($\chi^2 = 2.67, p = .102$, Cohen's $g = .33$) between the NVR and EVR object interaction task completion status.

Figure 6 (right) presents box plots of the perceived difficulty of the task for the NVR ($M = 2.92, SD = 1.68$) and EVR ($M = 3.42, SD = 1.44$) conditions. Wilcoxon Signed-Rank tests did not reveal a significant difference ($W = 33.5, p = .21, |r| = .49$) between the NVR and EVR conditions.

### D. Interaction Behaviors

Figure 7 plots the distribution of the number of EnVisionVR function activations for participants in the user study. The results show that the main objects indication function was activated a similar number of times for participants who regularly use assistive technology ($M = 2.00, SD = 1.20$) and for participants who do not ($M = 2.25, SD = 2.22$). However, participants who regularly use assistive technology activated the scene description function more ($M = 3.25, SD = 3.01$) than participants who do not regularly use assistive technology ($M = 1.25, SD = .500$). The object localization function was also activated more by participants who regularly use assistive technology ($M = 2.50, SD = 1.77$) compared with those who do not ($M = 1.25, SD = 1.26$). This suggests that participants with less visual perception capability tend to rely more on high-level scene descriptions and the fine detail object localization function. Meanwhile, participants with different vision capabilities relied on the Main Objects Indication Function at a similar level.

For participants who do not regularly use assistive technology, EnVisionVR appeared to complement their available vision. These participants used the scene description function less as they have enough residual vision to support their understanding of the scene, as evidenced by the performance of P2, P7, P9, and P11 in the scene understanding task in the NVR condition. They were also able to precisely locate small virtual objects as evidenced by successful completion of the object interaction task under the NVR condition. These participants used the main objects indication function more, likely because their residual vision does not allow them to explore a wide range in the scene, and they rely on the function to know what key objects are nearby.

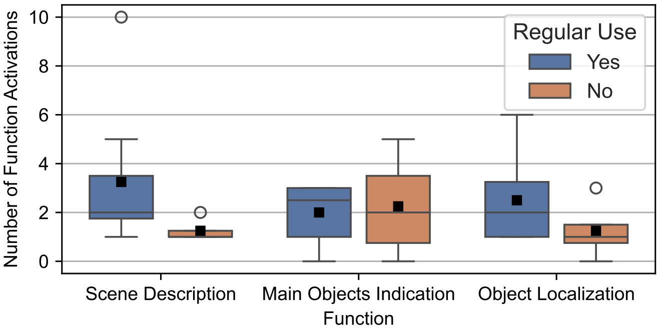


Fig. 7. Distribution of the number of EnVisionVR function activations for participants who regularly use assistive technology and for those who do not with labeled outliers. Black squares indicate the mean value.

### E. Post-Study Interview

In the post-study interview, 11 out of 12 participants expressed a preference for EnVisionVR over the NVR condition. Key themes were identified through a thematic analysis [18] and are summarized below.

*1) Level of Information Delivered:* The design of visual accessibility systems often faces a trade-off between the level of detail of information provided and the ease of use of the system. For the EnVisionVR system, participants liked how the scene description function was "helpful to identify a new location" (P2, P4, P5, P7, P10) with the "correct amount of detail" (P7) and "gave you a picture of the scene" (P3) and "a general overview" (P9, P10) to "build an image up in your mind" (P6). P11 commented that the tools could be more helpful if there was more detailed information. Overall, participants liked how the information had an appropriate amount of detail to gain a sense of physical presence.

> *"[With EnVisionVR,] I felt that I was in a real bar or a restaurant and I'd gotten out my camera [to get a description using Be My Eyes], it was very very good."* (P3)

Participants also liked how the main objects indication function told the user about key objects nearby. They liked how it told the user "if there is something on the right and left" (P2) and "helped users to be more confident by saying the names of things" (P5), but found the spatial tone to indicate the object location to be redundant (P4), and it sometimes did not pick up objects near the user (P9).

Participants found the object localization function helpful in providing the precise location of individual objects. Participants found it helpful in "telling [participants] how far [they] need to move" (P2), locating the object and letting the user "know the object is there" (P5), and "helping [participants] understand the direction of the object" and "gives good feedback" (P8). P4 found the beeping helpful in telling whether the controller was getting closer to the object. Overall, participants found EnVisionVR helpful in delivering both high-level scene information and detailed object-level information.

> *"[With EnVisionVR,] this is the first time that I have been able to do anything in VR. This is really promising, I think you're on to something here."* (P1)

*2) System Latency:* Participants commented that existing vision accessibility systems in the physical world, such as Be My AI, take several seconds to return a description. Participants contrasted this with their experience of EnVisionVR, which they found to be very responsive. P8 commented how she appreciated systems which give constant feedback on updates of what is in front of the user as the user moves.

> *"The advantage here is that the answer comes instantly and doesn't take any time to process the question."* (P3)

*3) Consistency in System Design:* The user study revealed that inconsistency in the system design could pose usability barriers. For example, the command "Where am I?" describes the user's field of view, while the command "What is near me?" reads out the names of objects which are near the user but not necessarily in front of the user. The inconsistency in reference frames sometimes led to confusion (P1, P9).

> *"Visually I could see where things were and I could move towards easily. The voice assistant wasn't giving me the instruction I needed, it didn't quite work. When I was looking for that 'brew button', it said it was next to me but I couldn't see it."* (P9)

As the scene descriptions were pre-baked based on images of the scene, there can be different names for the same object in the scene description function and the main objects indication function. For example, the bookholder on the desk in Anchor 1 of the test scene was referred to as a computer in scene descriptions, which can cause confusion for users.

*4) User Agency:* Participants also commented on different aspects of user agency over the system. These included user control on what information to deliver, the speed of delivery, and the level of detail of delivered information.

> *"I'd like this experience to have a speedier read-out speed, close to 200%. That really should be customisable. It'd be nice to have two levels of description, detailed and then summary."* (P1)
>
> *"I'd like a mode where I could scan a room, turning in my chair, and keep hearing an updated description of what's in front of me."* (P1)
>
> *"For me maybe [the voice description] was a bit slow. If you are in a new environment you don't want it too fast."* (P6)

Participants also suggested that the system could support more voice commands to improve user agency. P7 commented that the system was helpful but required users to memorize the different speech commands for each function. P12 reported difficulty in trying to remember how to phrase the question.

> *"The only thing I would say is maybe broaden the wording used to launch the command... If it was a bit more open in terms of voice commands."* (P6)

## VII. Design Implications

Results from our evaluation study reveal important design implications for VLM-assisted interactive systems for visual accessibility design in VR. Through these guidelines, we intend to assist designers and developers in creating more inclusive immersive systems for BLV users.

*1) Hierarchy of Descriptions:* Implement a hierarchy of high-level scene descriptions and detailed object level information. Results show that participants with regular use of assistive technology triggered scene descriptions more often (Section VI-D). 9 out of 12 participants appreciated the level of information delivered in scene descriptions, and 4 out of 12 users found the object localization function helpful (Section VI-E1). A 41.7% improvement in object localization is made with EnVisionVR (Section VI-B). These findings align with the observed benefits of adaptive levels of detail in scene descriptions [7].

*2) Dynamic Updates vs. Latency Requirements:* Adopt anchor-based VLM descriptions for predictable static scenes, and use dynamic on-demand updates of altered scenes and individual elements to account for object movement and user interaction. Participants favored EnVisionVR which balanced latency with flexibility (Section VI-E2). This design extends VLM-based systems like VR-GPT [26] by providing spatial information of objects in addition to 2D visual information in the user's field of view with low latency.

*3) Multimodal Feedback Consistency:* Standardize spatial reference frames (first- or third-person views) in speech descriptions and other output modalities. The evaluative study (Section VI-C and VI-E3) revealed inconsistent feedback as a cause of confusion. This extends prior work on audio description guidelines [4] to VLM-enhanced VR contexts.

*4) Customized Descriptions:* Align with existing assistive technology workflows to support customized reading speed and description verbosity. Blind users may want more detailed descriptions than low-vision users. The formative study (Section III-B3 and III-B4) showed strong user preferences for audio descriptions. Post-study interviews (Section VI-E1, VI-E4) also showed strong preferences for the right amount of detail in descriptions. This finding expands personalization in adaptive AR/VR [3] and VLMs [1] to customize VLM-generated descriptions in immersive environments.

*5) Feedback and User Agency:* Design flexible speech interfaces to support a wide range of commands while providing rich multimodal feedback. Participants suggested increasing the number of supported commands and overall control (Section VI-E4). This recommendation fits with the principles of Human-Centered AI [36] and the design principles (DP1 and DP2) on redundancy proposed by Dudley et al. [15].

## VIII. Discussion

EnVisionVR represents an original integration of high-level natural language scene descriptions and detailed object-level speech, audio, and haptic cues for object localization and interaction. We complement previous work on visual accessibility design in VR by incorporating VLMs to provide detailed scene descriptions to extend works such as SeeingVR [47] and VRBubble [22] which convert visual information to speech and audio, while also following Canetroller [46] and VIVR [25] in incorporating different feedback modalities to convey visual information such as the presence of a virtual object. We also demonstrate how it is possible to leverage speech, audio, and haptic information to design a multimodal system for VR visual accessibility design. Results from the user study show good promise in terms of supporting BLV users to enjoy VR experiences with the greatest benefit seemingly afforded to blind users or users with less usable vision.

The study results also reveal how EnVisionVR could be further improved. As EnVisionVR is intended to provide a proof-of-concept of how VLMs can be applied with other interaction modalities for visual accessibility design for VR content, the scene descriptions are pre-baked. This limits the current approach to static VR scenes. Future design iterations will aim to provide scene descriptions for dynamic VR scenes while balancing system latency. The evaluative study found different participants had different preferences in the verbosity and level of detail of scene descriptions. These examples demonstrate the significance of incorporating the ability to customize features for individual preferences, as well as adaptations for each user as they become more accustomed to the system. Additionally, we acknowledge that EnVisionVR is primarily a speech-driven interface with a limited number of supported commands. As is typical in accessibility research, we acknowledge the limitation of the small participant sample size and this must be considered when interpreting the findings. Our evaluation with 12 participants is comparable to that of SeeingVR's [47] 11 participants. It is also important to recognize that in both our work and SeeingVR [47], the sample includes participants representing a spectrum of visual acuity and vision loss. Given the need to prioritize time and reduce cognitive burden, we made a conscious decision to not ask BLV participants to reflect on task load and simulator sickness. Future design iterations of EnVisionVR will allow users to access more object and scene-level information through alternative and complementary forms of interaction.

## IX. Conclusion

This paper presents EnVisionVR, a proof-of-concept visual accessibility tool for VR based on scene descriptions and object-level guidance powered by VLMs, speech and audio cues, and haptic feedback. Our evaluation study with 12 BLV participants demonstrates the effectiveness of EnVisionVR in assisting scene understanding, object localization (41.7% increase in task success rate), and object interaction (33.3% increase in task success rate) for BLV users compared with the condition without visual accessibility features. We also summarize a list of design implications covering five different aspects of visual accessibility. We hope these findings and contributions will advance research in this space and ultimately lead to more inclusive VR experiences.

## Supplemental Material

The online appendix is available at https://osf.io/zb2ak/.

## Acknowledgments

Junlong Chen is supported by the China Scholarship Council and Cambridge Trust. This work is also supported by the Engineering and Physical Sciences Research Council (EPSRC), through the following grants: *Inclusive Immersion: Inclusive Design of Immersive Content* (EP/S027637/1 and EP/S027432/1) and *Towards an Equitable Social VR* (EP/W025698/1 and EP/W02456X/1). The authors thank Open Inclusion for their help in recruiting the research participants and administering the user research.

## References

[1] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. MyVLM: Personalizing VLMs for User-Specific Queries. In *European Conference on Computer Vision*, pages 73–91. Springer, 2024.

[2] Craig Anderton, Chris Creed, Sayan Sarcar, and Arthur Theil. Asleep at the virtual wheel: The increasing inaccessibility of virtual reality applications. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2025.

[3] Pradipta Biswas, Pilar Orero, Manohar Swaminathan, Kavita Krishnaswamy, and Peter Robinson. Adaptive accessible AR/VR systems. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[4] Hansjörg Bittner. Audio description guidelines: A comparison. *New perspectives in translation*, 20:41–61, 2012.

[5] Pietro Bongini, Federico Becattini, and Alberto Del Bimbo. Is GPT-3 All You Need for Visual Question Answering in Cultural Heritage? In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 268–281. Springer, 2023.

[6] Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV Ramakrishnan. More than meets the eye: A survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, pages 1–10, 2010.

[7] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. World-Scribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2024.

[8] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying Vision-and-Language Tasks via Text Generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[9] Brendan A Ciccone, Shannon KT Bailey, and Joanna E Lewis. The next generation of virtual reality: recommendations for accessible and ergonomic design. *Ergonomics in Design*, 31(2):24–27, 2023.

[10] Angelo Coronado, Sergio T Carvalho, and Luciana Berretta. See Through My Eyes: Using Multimodal Large Language Model for Describing Rendered Environments to Blind People. In *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences*, pages 451–457, 2025.

[11] Chris Creed, Maadh Al-Kalbani, Arthur Theil, Sayan Sarcar, and Ian Williams. Inclusive AR/VR: accessibility barriers for immersive technologies. *Universal Access in the Information Society*, 23(1):59–73, 2024.

[12] Khang Dang, Hamdi Korreshi, Yasir Iqbal, and Sooyeon Lee. Opportunities for Accessible Virtual Reality Design for Immersive Musical Performances for Blind and Low-Vision People. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, pages 1–21, 2023.

[13] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. LLMR: Real-time prompting of interactive worlds using large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024.

[14] Nicoletta Di Blas, Paolo Paolini, Marco Speroni, et al. "Usable Accessibility" to the Web for Blind Users. In *Proceedings of 8th ERCIM Workshop: User Interfaces for All, Vienna*, 2004.

[15] John Dudley, Lulu Yin, Vanja Garaj, and Per Ola Kristensson. Inclusive Immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse. *Virtual Reality*, 27(4):2989–3020, 2023.

[16] Be My Eyes, Sep 2023. Available at: https://www.bemyeyes.com/blog/announcing-be-my-ai. Accessed on December 4th 2024.

[17] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[18] Gregory Guest, Kathleen M MacQueen, and Emily E Namey. Introduction to Applied Thematic Analysis. *Applied Thematic Analysis*, 3(20):1–21, 2012.

[19] Jaylin Herskovitz, Jason Wu, Samuel White, Amy Pavel, Gabriel Reyes, Anhong Guo, and Jeffrey P Bigham. Making Mobile Augmented Reality Applications Accessible. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2020.

[20] Tina Iachini, Gennaro Ruggiero, and Francesco Ruotolo. Does blindness affect egocentric and allocentric frames of reference in small and large scale spaces? *Behavioural brain research*, 273:73–81, 2014.

[21] Open Inclusion. Home - Open Inclusion. Available at: https://openinclusion.com/. Accessed on Jan. 19th, 2025.

[22] Tiger F Ji, Brianna Cochran, and Yuhang Zhao. VR-Bubble: Enhancing peripheral awareness of avatars for people with visual impairments in social virtual reality. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–17, 2022.

[23] Lucy Jiang, Mahika Phutane, and Shiri Azenkot. Beyond Audio Description: Exploring 360° Video Accessibility with Blind and Low Vision Users Through Collaborative Creation. In *Proceedings of the 25th international ACM SIGACCESS conference on computers and accessibility*, pages 1–17, 2023.

[24] Claire Kearney-Volpe and Amy Hurst. Accessible Web Development: Opportunities to Improve the Education and Practice of Web Development with a Screen Reader. *ACM Trans. Access. Comput.*, 14(2), jul 2021. ISSN 1936-7228. doi: 10.1145/3458024.

[25] Jinmo Kim. VIVR: Presence of immersive interaction for visual impairment virtual reality. *IEEE Access*, 8: 196151–196159, 2020.

[26] Mikhail Konenkov, Artem Lykov, Daria Trinitatova, and Dzmitry Tsetserukou. VR-GPT: Visual Language Model for Intelligent Virtual Reality Applications. *arXiv preprint arXiv:2405.11537*, 2024.

[27] Masaki Kuribayashi, Kohei Uehara, Allan Wang, Shigeo Morishima, and Chieko Asakawa. WanderGuide: Indoor Map-less Robotic Guide for Exploration by Blind People. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2025.

[28] Diogo Lança, Manuel Piçarra, Inês Gonçalves, Uran Oh, André Rodrigues, and João Guerreiro. Speed-of-Light VR for Blind People: Conveying the Location of Arm-Reach Targets. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–5, 2024.

[29] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human Computer Interaction*. Morgan Kaufmann, 2017.

[30] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. VC-GPT: Visual Conditioned GPT for End-to-End Generative Vision-and-Language Pre-training. *arXiv preprint arXiv:2201.12723*, 2022.

[31] Sina Masnadi, Brian Williamson, Andrés N Vargas González, and Joseph J LaViola. 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, 2020.

[32] Microsoft. Seeing AI. https://www.microsoft.com/en-us/ai/seeing-ai, September 2021.

[33] Meredith Ringel Morris, Jazette Johnson, Cynthia L Bennett, and Edward Cutrell. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI conference on human factors in computing*

*systems*, pages 1–11, 2018.

[34] Vinaya Hanumant Naikar, Shwetha Subramanian, and Garreth W Tigwell. Accessibility Feature Implementation Within Free VR Experiences. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2024.

[35] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257, 2022.

[36] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.

[37] Kristina L Southwell and Jacquelyn Slater. An Evaluation of Finding Aid Accessibility for Screen Readers. *Information Technology and Libraries*, 32(3):34–46, 2013.

[38] Mauro Teófilo, Vicente F Lucena, Josiane Nascimento, Taynah Miyagawa, and Francimar Maciel. Evaluating accessibility features designed for virtual reality context. In *2018 IEEE international conference on consumer electronics (ICCE)*, pages 1–6. IEEE, 2018.

[39] Unity. VR Beginner: The Escape Room. https://assetstore.unity.com/packages/templates/tutorials/vr-beginner-the-escape-room-163264, October 2021.

[40] Yi Wang, Xiao Liu, Chetan Arora, John Grundy, and Thuong Hoang. Understanding vr accessibility practices of vr professionals. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2025.

[41] WebAim. Screen Reader User Survey #10 Results. https://webaim.org/projects/screenreadersurvey10/, 2024. [Online; accessed 13-August-2024].

[42] Kristin Williams, Taylor Clarke, Steve Gardiner, John Zimmerman, and Anthony Tomasic. Find and Seek: Assessing the Impact of Table Navigation on Information Look-up with a Screen Reader. *ACM Transactions on Accessible Computing (TACCESS)*, 12:1–23, 2019.

[43] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1180–1192, 2017.

[44] Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. VIAssist: Adapting Multi-Modal Large Language Models for Users with Visual Impairments. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*, pages 32–37. IEEE, 2024.

[45] Yizhe Zhang and Danny Z Chen. GPT4MIA: Utilizing Geneative Pre-trained Transformer (GPT-3) as A Plug-and-Play Transductive Model for Medical Image Analysis. *arXiv preprint arXiv:2302.08722*, 2023.

[46] Yuhang Zhao, Cynthia L Bennett, Hrvoje Benko, Edward Cutrell, Christian Holz, Meredith Ringel Morris, and Mike Sinclair. 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, 2018.

[47] Yuhang Zhao, Edward Cutrell, Christian Holz, Meredith Ringel Morris, Eyal Ofek, and Andrew D Wilson. 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, 2019.

[48] Jonathan Zong, Crystal Lee, Alan Lundgard, JiWoong Jang, Daniel Hajas, and Arvind Satyanarayan. Rich Screen Reader Experiences for Accessible Data Visualization. In *Computer Graphics Forum*, volume 41, pages 15–27. Wiley Online Library, 2022.

**Junlong Chen** is a PhD Student at the Department of Engineering, University of Cambridge. His research interests include scene interpretation for intelligent multimodal interactive systems and accessibility design.

**Rosella P. Galindo Esparza** is a Research Fellow at the Brunel Design School, Brunel University of London. She is a member of the Brunel Digital Design Lab, and her research focuses on human-computer interaction design for immersive technologies, accessibility, and social inclusion.

**Vanja Garaj** is a Professor of Design and the Director of Research at the Brunel Design School, Brunel University of London, where he also leads the Brunel Digital Design Lab, a research group specialising in design-led technology innovation. His research interests include human-computer interaction, accesibility and inclusive design.

**Per Ola Kristensson** is a Professor of Interactive Systems Engineering at the Department of Engineering, University of Cambridge and a Fellow of Trinity College, Cambridge. He is a co-founder and co-director of the Centre for Human-Inspired Artificial Intelligence (CHIA) at the University of Cambridge.

**John J. Dudley** is an Associate Teaching Professor of Machine Learning and Machine Intelligence at the Department of Engineering, University of Cambridge. He is a member of the Computational and Biological Learning Lab and his research focusses on the design of interactive systems that dynamically adapt to user needs and behaviours.