From Screen Reading to "Scene Reading" in SceneVR: Touch-Based Interaction Techniques for Use in Virtual Reality by Blind and Low-Vision Users

Melanie Jo Kneitmix
Paul G. Allen School of Computer Science & Engineering,
DUB Group
University of Washington
Seattle, Washington, USA
mekne@cs.washington.edu

Jacob O. Wobbrock
The Information School,
DUB Group
University of Washington
Seattle, Washington, USA
wobbrock@uw.edu

Figure 1: *SceneVR* is a touchscreen virtual reality (VR) controller that enables nonvisual access to virtual objects and avatars within a 3-D environment through a technique we introduce called "scene reading." *(Left)* The virtual scene is a campground, where objects are annotated with labels and descriptions for a touch-based "scene reader." *(Right)* The SceneVR touchscreen controller supports scene reading through touch gestures and spatial audio. SceneVR streams a real-time view from a VR headset to a user's phone and lets them "read" the scene with their finger to identify virtual objects and avatars that they touch, similar to techniques used in Slide Rule [38], Apple VoiceOver, or Android TalkBack, but for 3-D scenes rather than 2-D interfaces.

Abstract

To improve the accessibility of virtual reality (VR) for blind and low-vision (BLV) users, we introduce "scene reading," a technique inspired by touch-based screen reading for use in virtual environments. Scene reading provides semantic information about virtual objects and their on-screen positions, organizing details into hierarchies that users can navigate for more granular exploration; it also uses spatial audio for nonvisual feedback. To design and evaluate our scene reading technique, we developed a system called *SceneVR*, which supports touch and gesture input, and spatial audio output. SceneVR streams the live view from a VR headset onto a phone or tablet, letting BLV users drag their finger across the touchscreen to identify objects and avatars, navigate, and gain a spatial understanding of the scene. We conducted a task-based usability study to evaluate our SceneVR controller, collecting data on task performance, user experience, interaction patterns, and subjective feedback. Our findings indicate that scene reading with the SceneVR controller effectively supports BLV users in exploring virtual environments, enabling them to discover objects, navigate object hierarchies, and build an understanding of their surroundings while also providing a sense of enjoyment and agency. However, our findings also reveal initial design implications, including minimizing cognitive load and effectively integrating scene reading labels and descriptions with other sensory feedback to create a cohesive experience.

ASSETS '25, Denver, CO, USA

ACM ISBN 979-8-4007-0676-9/25/10
https://doi.org/10.1145/3663547.3746364

CCS Concepts

• **Human-centered computing → Accessibility technologies**; **Gestural input**; **Auditory feedback**.

Keywords

Accessibility, blind and low-vision, virtual reality (VR), touchscreen-based interfaces, spatial audio.

ACM Reference Format:

Melanie Jo Kneitmix and Jacob O. Wobbrock. 2025. From Screen Reading to "Scene Reading" in SceneVR: Touch-Based Interaction Techniques for Use in Virtual Reality by Blind and Low-Vision Users. In *The 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25), October 26–29, 2025, Denver, CO, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3663547.3746364

1 Introduction

Virtual reality (VR) is widespread, used for gaming [49], work [63], travel [68] and social interaction [32]. Despite its growing popularity, VR and its immersive 3-D digital content still pose significant accessibility barriers, especially for people who are blind or have low vision (BLV) and therefore cannot engage with its rich visual information [54]. For BLV users, basic tasks like identifying objects, understanding spatial relationships, navigating environments, or participating in social interactions can be difficult or impossible without alternative forms of access.

Screen reader technology has long been used to make 2-D digital content accessible to BLV individuals. Screen readers such as JAWS,[1] NVDA,[2] and Microsoft Narrator[3] provide nonvisual access to digital interfaces by converting displayed text into speech or braille output. They are typically controlled using a keyboard and mouse on desktop devices, or multi-touch gestures, like a tap, swipe, or drag, on mobile devices [4]. For visual elements such as images, screen readers use manually provided descriptions as alternative text (alt text) [31]. In addition to conveying content, screen readers can also communicate contextual information about interface elements. For example, some screen readers, like Apple VoiceOver[4] and Android TalkBack,[5] allow direct touch exploration to convey spatial layout or announce when text is part of a heading, menu, or button to indicate structural roles. These cues help users understand the screen's content and how it is arranged, enabling BLV users to form mental models of user interfaces.

Extending this kind of access to 3-D virtual environments, however, remains a challenge. VR lacks equivalent tools and standards for 3-D content despite calls from the community for screen reader-like solutions in commercial VR products [67]. Unlike static 2-D content, virtual environments are dynamic, often including complex objects with contextual relationships and interactive elements. Many objects in VR are agents, such as avatars, moving and interacting with others. Fully conveying this richness exceeds the capabilities of traditional screen readers, requiring new interaction techniques for 3-D exploration and scene understanding.

To address these challenges, we introduce "scene reading," a set of interaction techniques that enable nonvisual access to object-level annotations and their screen-relative positions in virtual environments [92]. Scene reading leverages hierarchical object information to progressively disclose detail systematically in response to users' touch and gesture interactions. To design and evaluate our scene reading techniques, we developed *SceneVR*, a touchscreen VR controller inspired by mobile screen readers like Slide Rule [38], Apple VoiceOver, and Android TalkBack, which facilitate screen reading through touch-based interactions, including direct touch exploration and multi-touch gestures. SceneVR streams the live view from a user's VR headset to the user's phone or tablet (see Figure 1). Users can explore the scene by scanning the screen with their finger. As they move their finger across the live view, SceneVR identifies objects and avatars they are touching to provide nonvisual access while supporting spatial awareness through direct touch and spatial audio.

To evaluate SceneVR, we conducted a task-based usability test [60] with 12 BLV adults to explore the perceived value and effectiveness of scene reading and SceneVR. Our findings indicate that scene reading with the SceneVR controller effectively supports nonvisual interaction, enabling object discovery and fostering a strong sense of presence while providing enjoyment and a sense of autonomy. However, our results also highlight initial design implications, including a steep learning curve associated with the touch-based interaction design and mismatches between user expectations and system feedback in multi-sensory environments. In particular, our findings show that annotations must work in tandem with sensory cues, as mismatches between what users perceive and what is described can disrupt their understanding and experience of the environment.[6] Finally, we offer initial design implications and future research directions aimed at enhancing nonvisual exploration and understanding of virtual environments for BLV users.

The contributions of this work are:

(1) The concept of "scene reading," a nonvisual interaction paradigm for virtual environments that draws from mobile screen readers and incorporates hierarchical organization, spatial audio, and progressive disclosure to support structured exploration and understanding of complex 3-D scenes.
(2) The implementation of scene reading in *SceneVR*, a prototype system that uses multi-touch gestures on a smartphone and spatial audio from a VR headset to enable nonvisual access to virtual environments, demonstrating how touch-based interaction can serve as an accessible interface for scene reading.
(3) Empirical results from a usability study evaluating SceneVR, including performance metrics, user experience assessments, interaction patterns, and subjective feedback.

2 Related Work

First, we review prior work on using mobile devices in augmented reality (AR) and VR, focusing on their efficacy as controllers. Second, we discuss the accessibility of 2-D visual content, highlighting

[1] https://www.freedomscientific.com/products/software/jaws/
[2] https://www.nvaccess.org/
[3] https://www.microsoft.com/en-us/windows/tips/narrator
[4] https://support.apple.com/en-al/guide/iphone/iph3e2e2281/ios
[5] https://support.google.com/accessibility/android/answer/6006598

[6] The challenge of automatically authoring 3-D object annotations was deemed beyond the scope of this work, which focused on interaction techniques for object discovery and spatial understanding. Automatic object annotation is an important avenue for future work, which we discuss in that section.

research on touch-based access methods that inform and inspire our approach in 3-D virtual environments. Third, we review research on exploration and scene understanding for BLV VR users.

2.1 Mobile Touchscreens as VR Controllers

Prior work has explored the use of mobile devices as interfaces for AR and VR. Early research by Feiner et al. [21] and Szalavári et al. [80] envisioned systems that combined handheld technologies with head-mounted displays (HMDs), using multiple displays and interaction techniques to enhance mobility and ease of use. Since then, research has continued to use mobile devices with HMDs, focusing increasingly on how to integrate modern touchscreen devices. For example, Grubert et al. [27] developed a system that distributed widgets across an AR headset, smartphone and smartwatch, effectively minimizing the interaction seams across the growing number of devices that we use daily. Others have examined interactions between headsets and mobile touchscreens for knowledge-based tasks, highlighting the potential of these interfaces to enhance productivity and serve as effective input tools [9, 34, 44].

Highly relevant to our work is a growing body of research demonstrating the feasibility of using mobile touchscreen devices as input controllers for AR and VR HMDs, motivated in part by the ubiquity of these devices. Babic et al. [6] introduced Pocket6, a smartphone-based six-degrees-of-freedom (6DoF) controller using mobile AR tracking, finding its performance comparable to commercial 6DoF controllers. TrackCap [51] extended this approach by enabling inside-out tracking using the smartphone's camera, showing improved accuracy and task completion times over 3DoF controllers. BiSHARE [98] explored bidirectional interaction between smartphones and AR HMDs, including using the phone as an AR controller, and PAIR [87] provided 6DoF smartphone input for AR with a larger tracking volume compared to prior work. Phonetroller [50] and HandyCast [41] further demonstrated the potential of smartphones as VR controllers. Phonetroller visualized finger positions to support precise touchscreen interactions, while HandyCast enabled bimanual control of virtual hands through compact phone gestures in physically constrained spaces.

Beyond demonstrating feasibility, prior work also suggests that smartphone-based controllers may outperform conventional input devices in certain contexts. Touchscreen input has been shown to better support precise target acquisition tasks compared to ray casting with conventional controllers [47], and to significantly reduce error rates during text entry [11].

We build on prior work demonstrating the technical viability of mobile touchscreens as VR controllers for sighted users, shifting focus to their potential for nonvisual interaction. Although previous systems have not addressed accessibility, they establish mobile touchscreens as a feasible platform for our approach. We extend prior work by re-purposing a mobile phone as an accessible VR controller, leveraging its touchscreen as a device for nonvisual exploration.

2.2 Accessibility of Touchscreen-Based 2-D Visual Content

Touchscreen-based exploration of 2-D user interfaces (UIs) has been widely studied. Early research on touchscreen kiosks by Vanderheiden [88] offered the "talking fingertip technique," which announced buttons on an ATM when they were touched. Kane et al. significantly extended this idea to enable BLV access to smartphones, tablets, tabletops, and documents in Slide Rule [38], Access Overlays [40], and Access Lens [39]. Work by Goncu et al. [25] used vibrotactile feedback to make computer graphics more accessible.

More recently, touch-based exploration techniques have been applied to enable understanding and authoring of digital artboards [94, 95], and data visualizations [93], both of which are rich 2-D spaces that have been shown to be particularly inaccessible for BLV users [69, 73]. Similarly Sharif et al. [72] explored the use of a touchpad as an input device for screen readers, enabling users to spatially explore digital content and video elements mapped to the touchpad's surface.

Perhaps most extensive are studies of digital image accessibility. For example, Morris et al. [53] investigated several techniques to supplement alt text, including image segmentation, to enable touch-based exploration and tiered descriptions that disclose more detail as the user requests it. Building on this, Ahmetovic et al. [1] compared attribute-based segmentation with a hierarchical approach, where child components were revealed after their parent was explored. Although both had benefits, hierarchical exploration was more engaging to participants and was also associated with more detailed descriptions of the digital artwork they explored during the study. Lee et al. [45] investigated whether touch-based exploration helped BLV users assess the correctness of AI-generated captions. Their findings reinforce the benefits of touch-based techniques and hierarchical organization for image understanding, particularly in supporting spatial awareness, agency, and user control, but also highlighted trade-offs in effort and efficiency when compared to simpler, text-based approaches. To address usability issues while preserving the benefits of direct touch interaction, Nair et al. [57] developed ImageAssist, a system that introduced additional tools to support and scaffold touch-based image exploration. Their study found that participants appreciated the added tools and preferred using them as a complement, not a replacement, to free-form direct touch exploration.

These research-driven designs have also begun appearing in commercial products. For example, Microsoft's Seeing AI[7] introduced the *Explore* feature, which enables spatial exploration of segmented digital images on a phone. This real-world adoption highlights the growing relevance of touch-based exploration techniques, and collectively, these technologies mark a shift from static image descriptions to dynamic exploration.

Informed and inspired by this work, our research investigates whether touch-based exploration with progressive disclosure can be adapted to dynamic 3-D environments, similarly fostering a rich understanding of virtual scenes through knowledge of objects, their hierarchies, and their relative positions.

[7] https://www.seeingai.com/

2.3 Scene Understanding for BLV VR Users

Scene understanding is essential for effective interaction in virtual spaces. Franz et al. [23] reviewed literature from computer vision and cognitive psychology to develop a taxonomy for how users construct this understanding, highlighting the importance of objects and spatial awareness in forming a conceptual model of virtual scenes. Their work further highlights that exploration and the ability to perceive and interact with these objects are foundational to other virtual environment tasks [23]. For BLV users, scene exploration and understanding requires nonvisual access to this information, enabling them to discover objects, develop spatial awareness, and build a mental model of the environment.

Research on BLV VR accessibility has explored various methods of providing access to objects and their layout. Early work focused on free-form exploration using mouse and keyboard input to support object identification, navigation, and orientation in 3-D games like Second Life [15, 22, 43, 65, 66] and PowerUp [83, 84]. These studies demonstrated that providing object descriptions and spatial information can improve accessibility, particularly when interactions align with familiar screen reader techniques. However, these methods also degraded interaction speeds compared to visual alternatives, and their effectiveness in fully immersive VR environments remained unaddressed.

Later work adapted object description techniques for modern VR headsets and commercial controllers. However, this research found that BLV users struggled with traditional selection methods such as ray casting, which relies on pointing-based interactions that are challenging without visual feedback [20]. SeeingVR [97] also incorporated object description techniques for modern VR applications but as part of a broader suite of tools aimed at improving accessibility for low-vision users. Chen et al. [16] explored the use of Vision Language Models (VLMs) for scene interpretation in VR, developing EnVisionVR, a system that processes headset-captured images to generate nonvisual scene descriptions and support object localization through multi-modal feedback. Meanwhile, research by Nair et al. [55, 56, 58] investigated free-form exploration in video games, developing techniques that let BLV players enrich their understanding of their surroundings through discovery-driven navigation. Herskovitz et al. [33] addressed similar challenges in iOS AR by creating a bridge that exposes AR content to VoiceOver and developing guided and free-form interaction techniques to help blind users locate and explore virtual objects in physical space.

Beyond verbal object descriptions, research has also explored how multi-sensory feedback, particularly auditory and haptic cues, can further support BLV access in VR. For example, Balasubramanian et al. [7] described the Scene Weaver prototype in which users could navigate virtual environments by exploring people, places, and objects through a self-directed interface, emphasizing the importance of supporting individual exploration strategies through perceptual agency. Some approaches have focused on refining auditory representations of virtual objects and their layout, including research into sonification techniques [82], the design space of common auditory feedback methods [28], and echolocation [5, 81]. Others have explored haptic feedback as an additional sensory cue to enhance object and environmental awareness [18, 42, 79, 86]. Additionally, haptic feedback has been incorporated into social interactions [17, 37] and games [19, 24, 52, 91] to improve accessibility through multi-sensory feedback. Yet others have incorporated tools like canes [48, 74, 85, 96] and gloves [26, 62] to enable exploration through more tactile experiences.

More recently, commercial devices have begun to incorporate accessibility features for BLV users. Notably, the Apple Vision Pro includes support for VoiceOver,[8] which can render menus, system interfaces, and other digital content in the AR environment nonvisually, making it one of the first commercially available AR headsets to provide built-in screen reader access to elements of the immersive interface.

Building on this rich set of work, we present the first exploration of touch-based "scene reading" in VR, introducing and evaluating finger-driven techniques that enable BLV users to access, explore, and understand virtual environments through progressive disclosure of details via VR object hierarchies.

3 "Scene Reading" and the Design of *SceneVR*

To explore and evaluate scene reading, we designed *SceneVR*, a native iOS (or iPadOS) app that serves as a touchscreen VR controller and instantiates scene reading using touch-based interactions. SceneVR uses multi-touch gestures to support nonvisual exploration and understanding of virtual environments. This section describes both the broader design principles of the concept of scene reading and its particular touch-based implementation in SceneVR, including the touch-based input gestures shown in Figure 4.

3.1 Scene Reading Gestures

Scene reading supports nonvisual exploration and understanding of virtual environments by providing access to objects and their on-screen layout. Unlike traditional screen readers, which operate within a 2-D coordinate system, scene reading supports interaction within 3-D space, where objects exist within dynamic and interactive environments. To navigate this added complexity, scene reading incorporates features such as free-form and structured exploration, and organizes objects into natural hierarchies that users can navigate to progressively reveal additional detail.

In our touch-based SceneVR controller, scene reading is enabled mostly through one-finger touch gestures. We adapt touch-based techniques designed for nonvisual access to images [1, 45, 46, 53, 57], graphics [25, 93–95] and user interfaces [38–40]. We also draw on interaction patterns used in mobile screen readers like VoiceOver and TalkBack, which rely on multi-touch gestures to support interface navigation and access.

One-finger drag for spatial scene reading. A recurring theme in prior work is the role of direct touch in enabling free-form exploration of visual content while also supporting spatial awareness through continuous movement and audio or haptic feedback [1, 25, 38–40, 45, 46, 53, 57, 88, 93–95]. This form of interaction is also supported in VoiceOver and TalkBack, which use drag gestures to let users hear the on-screen element beneath their

[8] https://support.apple.com/en-al/guide/apple-vision-pro/tanae5174040/visionos

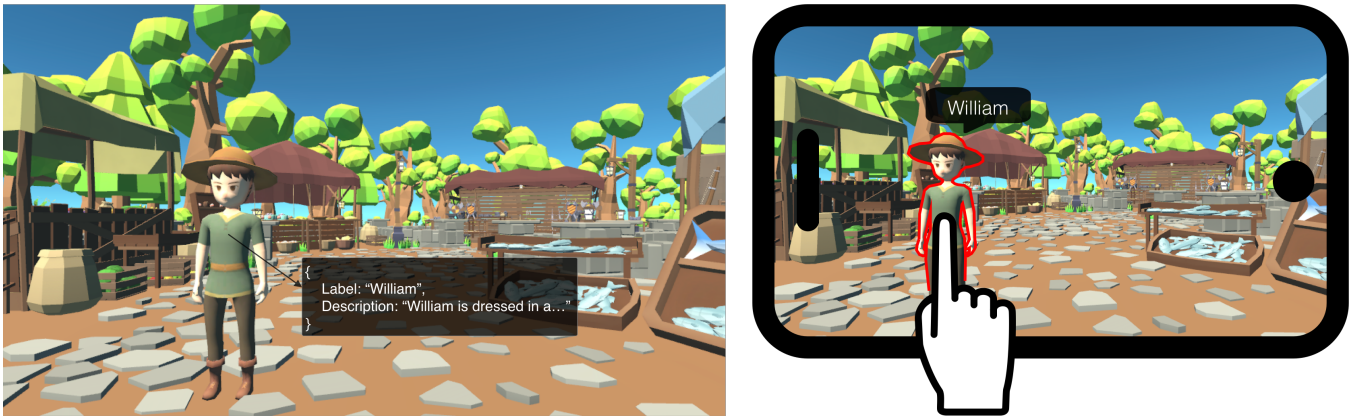


Figure 2: SceneVR enables spatial scene reading through direct touch. *(Left)* A virtual marketplace as viewed from a VR headset, where an avatar is annotated with a label and description. *(Right)* The same view is streamed to and displayed on a phone by our SceneVR controller, where the avatar is selected and its label announced as the user's finger drags over it. Additional gestures can access further information (described below).

touch point. Leveraging insights from this work, we designed spatial scene reading in SceneVR to apply direct touch exploration to 3-D virtual environments.

To support spatial scene reading, SceneVR streams the live view from a VR headset to a touchscreen display and enables exploration using a one-finger pan (Figure 4a). To keep scene reading consistent and predictable, the live-streamed view does not rotate with the user's head, which may shift due to small, unintentional movements. Instead, the view is only rotated when the user intentionally turns their avatar's forward direction. (This rotation interaction is described in more detail below.) As users move a finger across the view, SceneVR selects the object they are touching by translating the touch point from screen coordinates to a 3-D ray cast into the virtual world[9], identifying the first object the ray intersects. As the user's touch moves across an object, its label is announced using spatial audio (e.g., "*Vegetable Stall*"), which also conveys its relative position in the virtual space. Figure 2 illustrates this process.

Split-tap for detailed description. Similar to the approach used in the Slide Rule finger-driven screen reader to acquire targets [38], and later adopted by VoiceOver, TalkBack, and A11yboard [95], we also use a split-tap gesture whereby a second finger tap anywhere on the screen triggers a more detailed audio description of the object currently being touched by the "reading finger" (Figure 4b). These descriptions include visual details similar to traditional alt text; they can also provide information relevant to the virtual nature of the object, such as the presence of a nearby teleport location, which allows users to jump directly to that object (e.g., "*A wooden stall displays baskets of tomatoes, carrots, and leafy greens. The stall is shaded by an awning. Teleport available.*").

One-finger flick left/right for sequential scene reading. Prior work in touch-based image exploration [57] and BLV video game [55] and AR [33] accessibility shows that users benefit from both free-form and structured exploration methods, where structured methods present objects in a predefined, predictable order. VoiceOver and TalkBack similarly support both modes: users can explore freely by dragging their finger or navigate in-order using a flick gesture. To examine whether structured exploration provides similar benefits in 3-D environments, we also introduce sequential scene reading, where users can move through objects in a predefined order. SceneVR implements this technique using a one-finger flick gesture (Figure 4c).

Unlike spatial scene reading with a continuously moving finger, discrete flick gestures do not depend on touch location. Instead, a one-finger flick right moves to the next object clockwise, while a one-finger flick left moves counterclockwise, with virtual objects ordered according to their radial position around the user, providing for egocentric orientation. As with spatial scene reading, an object's label is announced using spatial audio and the object remains selected until the user performs another scene reading gesture.

One-finger circle for overview scene reading. Prior work in touch-based image exploration also highlights the value of providing an overview of visual content along with the ability to examine individual elements [45, 57]. VoiceOver and TalkBack offer similar functionality via a gesture that announces all elements in sequence. While SceneVR does not replicate the exact touch gesture used in either screen reader, we introduce a similar capability in 3-D environments through overview scene reading, which allows users to hear the names of all visible objects. SceneVR implements this technique using a one-finger circle gesture (Figure 4d).

Unlike spatial (continuous, direct touch) or sequential (discrete flicks) scene reading, which reveal individual objects, overview reading provides a high-level scene summary. When users trace a complete circle with one finger on the SceneVR touchscreen, the system announces the labels of all currently visible objects in a predefined left-to-right order, based on the user's view. Because all objects are announced in a consistent manner and order, overview scene reading also serves as a second form of structured exploration, along with sequential scene reading. As with other scene reading techniques, spatial audio conveys approximate object locations.

[9]https://docs.unity3d.com/6000.0/Documentation/ScriptReference/Camera.ScreenPointToRay

Hierarchical progressive disclosure. Prior work in touch-based image exploration has used progressive disclosure [61] to present detailed information in a more manageable way for users. Strategies include presenting alt text descriptions incrementally, allowing users to request additional details as needed [53] and structuring images into object hierarchies that users can navigate for more granular exploration [1, 45, 46, 57].

Dense VR environments with many objects can make nonvisual exploration cognitively overwhelming. To explore whether progressive disclosure alleviates this challenge, we introduce two hierarchy-based methods for revealing detail: (1) user proximity, which automatically reveals greater Level of Detail (LOD) with greater user avatar proximity, and (2) object groups, which require user interaction to access greater LOD. For user proximity-based disclosure, child objects are revealed as the user's avatar approaches the parent, similar to how visual elements are progressively rendered in game design [90]. For example, as a user explores a virtual market using scene reading, they might first hear a label for a distant "*Bakery Stall*." As their avatar moves closer, additional annotations are revealed, describing individual goods and displays *within* the stall. This form of disclosure is illustrated in Figure 3.

One-finger flick up/down to navigate object groups. Although increasing LOD based on user proximity to objects reveals increasing detail as the user moves, proximity alone may not always be practical or achievable. Some objects, like a drink machine behind a service counter in a virtual restaurant, cannot be physically approached. In other cases, small or cluttered objects require precise camera positioning, making spatial scene reading difficult. To address these limitations, we introduce *object groups*. Inspired by item groups in Apple's VoiceOver,[10] object groups are an alternative method of progressively navigating object hierarchies and are designed to reduce the burden of manual positioning.

SceneVR implements object-group navigation with a one-finger flick up/down gesture (Figure 4c). After selecting an object with an associated group, users can enter the group with a one-finger flick up, temporarily repositioning the virtual camera to a predefined optimal viewpoint that centers on the details within the group. When inside a group, unrelated objects are filtered from scene reading, letting users focus on details without distraction. Users can exit the group at any time with a one-finger flick down, restoring the previous scene-reading view. For example, a virtual drink machine can define an object group that contains its dispensers. After selecting the machine, a one-finger flick up shifts the virtual camera to center on the drink dispensers and limits scene reading to only those objects. When finished, a one-finger flick down restores the user's previous view and lets them resume exploration of the broader scene.

3.2 Locomotion Gestures

Since our focus is on scene exploration and understanding, we designed SceneVR primarily for object discovery rather than to replicate the full range of VR interactions found in commercial controllers. However, we also incorporated basic locomotion, since movement is fundamental to scene exploration. Unlike scene reading, which relies mostly on one-finger gestures, locomotion uses two-finger gestures to turn, walk, and teleport.

Two-finger drag and hold left/right to rotate. Rotating is performed with a two-finger drag and hold gesture, where users place two fingers anywhere on the touchscreen, drag left or right, and hold to turn in that direction (Figure 4e). Rotation continues while two fingers remain on the screen and stops when they are lifted. A continuous audio cue, anchored to the user's original forward direction, indicates rotation progress. For example, when turning right, the cue will be heard from the left ear at 90 degrees. To aid in spatial awareness, the system also announces object labels as they come into view, integrating scene reading feedback during the turning process. Rather than calling out every object that enters the field of view, the system announces objects only when they are centered in the user's egocentric perspective. When the user lifts their fingers and stops turning, the continuous audio cue fades, and their new forward direction is conveyed though speech output (e.g., "*Facing west*").

Two-finger drag and hold up/down to walk. Similar to rotation, walking is also initiated by a two-finger drag and hold gesture, but with an upward drag moving the user forward and a downward drag moving the user backward without turning (Figure 4e). Movement continues while the fingers remain on the screen and stops when they are lifted. Footstep audio plays while the user is walking, and upon stopping, the system announces the distance and direction moved (e.g., "*2 meters forward*").

Two-finger tap to teleport. Users can teleport to specific locations within the virtual environment using a two-finger tap (Figure 4f). After selecting an object with a defined teleport location (specified via object attributes in Table 1), a two-finger tap jumps the user to that location. Upon arrival, the system announces the new position using the selected object's label (e.g., "*At Vegetable Stall*").

4 Study Method

To evaluate our scene reading techniques and SceneVR prototype for enabling BLV exploration and understanding of virtual environments, we conducted a task-based usability study [60]. In this section, we describe our study design.

4.1 Participants

We recruited one low-vision and 11 legally blind adults from the local area using community organizations and mailing lists. Participant age ranged from 32 to 75 years old ($M = 51.82$, $SD = 13.50$). Six participants identified as men and six as women. All participants had basic proficiency using touchscreens, including screen readers, and could perform multi-touch gestures, like a one- or two-finger drag or tap, on a touchscreen device. Six participants had prior experience with VR headsets. Of these, five described the systems they used as difficult to operate or understand, while one reported no major issues but noted having more vision at the time of use.

[10] https://support.apple.com/en-au/guide/iphone/iphfa3d32c50/ios

Figure 3: SceneVR uses hierarchical object organization to progressively reveal detail during scene reading. Disclosure is governed by either (1) user proximity, where greater Level of Detail (LOD) is revealed with greater proximity, or (2) object groups, which let users manually explore detail belonging to a parent object. This example demonstrates the first LOD disclosure, where (*left*) from a distance, scene reading presents the "Bakery Stall" as a single item. As the user moves closer (*right*), SceneVR progressively reveals individual elements within the stall, such as tables, displays, and baskets.

4.2 Apparatus

We built a single Unity app, running on a Meta Quest 2 headset, that contained three virtual environments: a simple tutorial environment (Figure 5a) and two more detailed test environments for the task-based usability study (Figures 5b and 5c). We built these environments using assets from the Unity Asset Store [2, 10, 59, 64, 77, 78, 89]. The SceneVR controller, a native iOS (or iPadOS) app, ran on an iPad mini 6 and facilitated interaction with the Unity app using low-latency peer-to-peer communication via the WebRTC protocol.[11] The WebRTC protocol was implemented through a Unity package[12] and an open source iOS library [76].

The WebRTC session established a data channel for communication, allowing the SceneVR controller to notify the Unity app when an input gesture was recognized, as well as a video stream to display the Quest headset's view on the SceneVR controller. Although BLV SceneVR users did not rely on the video stream while wearing the headset, it was essential for debugging and provided us with visual confirmation of direct touch interactions and system responses.

The SceneVR controller primarily relied on UIKit gesture recognition[13] to recognize user input, with one exception: the circle gesture (Figure 4d) was detected using the OpenCV library [13] to fit a circle to a collection of touch locations.

For audio feedback, the Unity app used 3-D audio sources from Unity's scripting API,[14] enabling spatial audio during exploration. Text-to-speech (TTS) output was generated using Azure's TTS services.[15]

Figure 6 shows a study participant wearing the Meta Quest 2 headset and interacting with the virtual environment using the SceneVR controller.

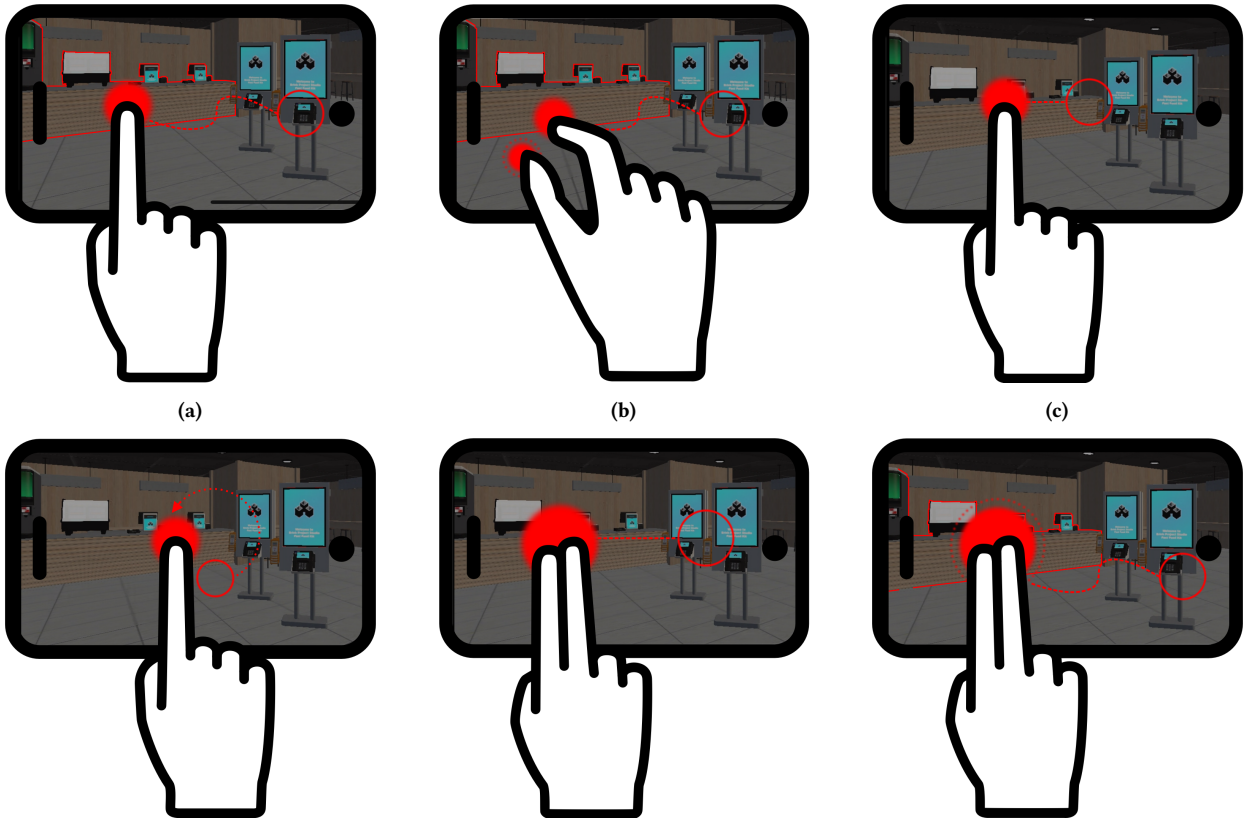
[11] https://webrtc.org/
[12] https://docs.unity3d.com/Packages/com.unity.webrtc@2.4
[13] https://developer.apple.com/documentation/uikit/handling-uikit-gestures
[14] https://docs.unity3d.com/ScriptReference/AudioSource
[15] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/text-to-speech

Unity Game Object Attributes. We created a set of attributes, listed in Table 1, that can be added to any game object in a Unity VR app to make it accessible to our scene reader. Although we did not evaluate this attribute set in our study, we were mindful of the need for future work to explore how such accessibility information could be standardized, automated, and scaled. We discuss directions for this in future work. For our study, we assumed a well-annotated environment, with the research team manually assigning the necessary attributes to the objects and avatars in each scene, enabling us to design, iterate, and evaluate scene reading in SceneVR.

4.3 Procedure

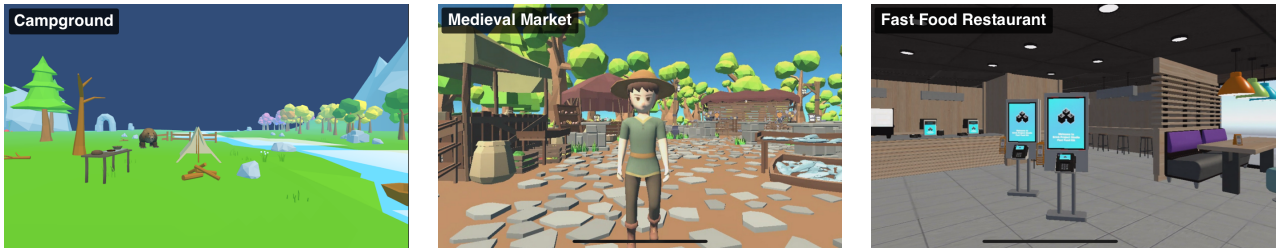
We conducted the study on the University of Washington campus and in the greater Seattle area. The study consisted of four parts: (1) a pre-study interview to gather demographic information and identify prior experience with and expectations of VR technology, (2) a 15 minute tutorial on the SceneVR controller and its scene reading capabilities, (3) a task-based usability study where we asked participants to complete exploration tasks using SceneVR, and (4) a post-study interview to gather participants' feedback on their experience with the system. All parts of this research occurred during a single one-hour session with each participant. This study was designed to evaluate the SceneVR system in depth and did not include comparative conditions. We discuss the reasoning behind this decision, including the lack of existing comparisons for BLV VR users, in Section 7.

Tutorial. Before the task-based assessment, participants could adjust the Apple iPad according to their desired accessibility settings, and adjust the volume and fit of the Meta Quest 2 headset. The first author then introduced the tutorial environment and provided instructions on using the SceneVR controller and its scene reading interactions. Afterwards, participants could continue to practice until they felt comfortable to begin the task-based assessment. On average, the tutorial lasted about 14.19 minutes (SD = 3.35) per participant.



(a) (b) (c) (d) (e) (f)

Figure 4: Multi-touch gestures for exploring virtual environments using SceneVR, in this case, a fast food restaurant: (a) one-finger drag (placing a finger on the screen and scanning) for spatial scene reading to identify objects being touched, (b) split-tap (tapping anywhere with a second finger while still holding the first finger down) to receive the description of the first finger's object, (c) one-finger flick (quickly swiping a finger across the screen) for sequential reading to move through objects in order (*left/right*) or to move in and out of object groups (*up/down*), (d) one-finger circle (tracing a circular motion with one finger) for overview reading to announce all objects currently in view, (e) two-finger drag and hold (placing two fingers on the screen, dragging, and holding in place) to rotate (*left/right*) or to walk forward and backward (*up/down*), and (f) two-finger tap (tapping two fingers simultaneously on the screen) to teleport to the selected object.



(a) (b) (c)

Figure 5: The virtual environments used in our usability study: (a) a campground used for the system tutorial, (b) an outdoor medieval market with five stalls, and (c) an indoor fast-food restaurant.

Attribute	Data Type	Description	Example
Label*	String	Short identifier	*Vegetable Stall*
Description*	String	Detailed description, including visual features and other relevant virtual attributes.	*A wooden stall displays baskets of tomatoes, carrots, and leafy greens. The stall is shaded by an awning. Teleport available.*
Progressive Disclosure Method	Enum { *User Proximity*, *Object Group* }	Governs how child objects are disclosed. For user-proximity disclosure, child objects are revealed when the user is in close proximity. For object-group disclosure, child objects are disclosed and moved into focus when the user inspects the group.	Items sold at the *Vegetable Stall* are revealed as the user approaches (user proximity). The dispensers available at the *Drink Machine* are revealed and brought into focus when the user inspects the *Drink Machine* group (object group).
Parent	Unity Game Object	The parent object within the hierarchy. An object with a *Parent* will be disclosed according to the parent's *Progressive Disclosure Method.*	The parent of an iced tea dispenser is the *Drink Machine.*
Focus Position* for *Object Group* disclosure	3-D Vector	The predefined camera position for focused scene reading in an object group, centering on child objects for easier exploration.	After entering the *Drink Machine* group, the camera temporarily repositions for a focused view of the available dispensers.
Teleport Position	3-D Vector	A teleport location near the object.	The user can teleport to the *Vegetable Stall.*

Table 1: Attributes the research team added to virtual game objects in our Unity test environments to enable scene reading. *An asterisk (*) marks required items.*

Figure 6: A BLV participant using the SceneVR controller to explore a virtual environment during our usability study. They are wearing a Meta Quest 2 headset, which runs the test environments used in the study, while interacting with the virtual scene via the SceneVR controller on an iPad Mini 6.

Task-Based Assessment. The task-based assessment used two virtual test environments, an outdoor medieval market and a modern-day fast-food restaurant, and participants completed similar tasks in each environment. To begin, participants were given three minutes to freely explore the virtual environment to further practice using SceneVR and become familiar with their virtual surroundings.

Following free exploration, participants completed structured tasks to evaluate SceneVR and its scene-reading capabilities. To ensure representative task scenarios, we followed guidance that recommended mixing simpler, atomic tasks and more complex tasks requiring higher-level cognitive processing [12]. We focused our tasks on scene exploration and understanding, and presented four structured tasks in each environment: two object-finding tasks (*Parent Object (PO)* and *Child Object (CO)* tasks) and two inference-based tasks requiring participants to draw conclusions from multiple objects (*Scene Context (SC)* and *Spatial Awareness (SA)* tasks):

- *Parent Object (PO)*: Locate a prominent object at the top level of the object hierarchy. For example, we asked participants to locate the meat stall in the medieval market and an eight-top table in the fast food restaurant.
- *Child Object (CO)*: Locate an object nested in a parent by navigating a hierarchical structure governed by both proximity-based and object-group disclosure. In the medieval market, we asked participants to choose a dessert from a display case at the bakery stall; in the fast-food restaurant, we asked them to select a drink from the drink machine located behind the service counter.
- *Scene Context (SC)*: Infer context based on nearby objects. In the medieval market, we asked participants to identify a stall based on its items; in the fast food restaurant, we asked them to infer the restaurant type from the food on a table.
- *Spatial Awareness (SA)*: Assess configuration knowledge, i.e., knowledge of how objects are located in relation to one another in the virtual environment [8]. We asked participants to describe the on-screen location of one landmark relative to another in each environment.

To control for order effects, we counterbalanced the order of the virtual environments and structured tasks. We also imposed a three-minute time limit for each structured task to ensure participants had ample opportunity to attempt all tasks in the assessment. No participant exceeded the time limit on any task. If at any point participants forgot how to use the SceneVR controller or its gestures, they were allowed to ask the researcher for assistance. (The frequency of such requests is discussed in the results section.)

Post-Study Interview. After participants completed the task-based usability test, we asked them a series of Likert [36] and open-ended questions to better understand their experience of scene reading and SceneVR. Likert questions included the NASA Task Load Index (NASA-TLX) [29, 30] and iGroup Presence Questionnaire (IPQ) [70, 71] to assess perceived workload and sense of presence, respectively. We modified the IPQ to remove vision-related questions.

4.4 Data Analysis

SceneVR usage metrics were logged on-device, and the research team recorded task performance and interview responses in a digital spreadsheet. With participant permission, we also recorded and transcribed session audio. Quantitative analysis included descriptive statistics for task performance, user satisfaction, NASA-TLX and IPQ scores, as well as an examination of usage patterns related to scene reading interactions. For qualitative analysis, we applied inductive coding [14] to identify themes in participants' open-ended responses. As part of this process, we used affinity diagramming [35] to organize insights and develop inductive codes from transcript data. One researcher conducted an initial round of coding, followed by a peer debriefing process [75] in which another researcher used the code book to analyze two transcripts. The researchers then met to discuss and resolve discrepancies, reaching the final code book.

5 Results

We present the results of our usability study, examining task performance, user experience, interaction patterns, and qualitative feedback. We take each of these in turn, below.

5.1 Task Performance and User Satisfaction

Eight of 12 participants successfully completed the assigned tasks in the task-based assessment; four participants failed a spatial awareness (SA) task in one of the two test environments. Additionally, one of these participants also failed the scene context (SC) task in one of the test environments. Table 2 presents each participant's performance across tasks in both environments. Of 96 total tasks, participants were successful in 91 of them (94.8%).

Following the task-based assessment, participants rated their satisfaction with SceneVR on a 7-point Likert scale [36], with 7 being the highest, reporting overall high satisfaction ($M = 5.92$, $SD = 1.24$). Figure 7 presents a histogram of participants' satisfaction responses. Qualitative feedback reinforced these results, with many participants (P01, P03, P04, P05, P07, P09, P10, P11) expressing enthusiasm for the system. For example, P07 remarked, "I think this is pretty cool," while P11 stated, "This is incredible. Well, I wish I could take it home." Although this excitement may reflect initial novelty, eight of 12 participants (P03, P04, P05, P06, P07, P08, P11, P12) specifically described SceneVR as fun and enjoyable to use, suggesting engagement beyond first impressions. Additionally, four of 12 participants (P02, P04, P08, P12) highlighted its value in fostering a sense of independence, emphasizing its ability to be used without assistance. For example, P04, who had prior VR experience, noted, "When I did this [VR] with my son, he had to show me. This [SceneVR] I could do independently."

Figure 7: Distribution of participant satisfaction with SceneVR, where participants rated their satisfaction on a 1-7 scale, with 7 indicating high satisfaction. Overall, participants reported positive feedback ($M = 5.92$, $SD = 1.24$).

Participants' qualitative feedback also highlighted specific features of SceneVR they found useful. Here, we present feedback on object-group navigation, which was particularly relevant to the child object (CO) task; this task asked participants to inspect an object group to locate a nested object. In the post-study interview, two participants (P04, P08) specifically mentioned the object-group feature as being especially useful. P04 described object-group navigation as intuitive, noting that "the ease and ability to move around or to get into a group and look at the group" was something they particularly liked about using SceneVR. Similarly, P08 found object groups helpful, stating, "The group search function was actually really useful," and later elaborating that the groups were especially beneficial when examining "very fine details."

5.2 Workload and Learning Curve

We used the NASA Task Load Index (NASA-TLX) [29, 30] to assess participants' perceived workload while using SceneVR, measuring workload across six dimensions: mental, physical, and temporal demand; performance; effort; and frustration. Participants rated each dimension on a 7-point scale, where lower scores indicate less demand, effort, or frustration, and better performance. Hence, lower is better on all scales.

Participants reported low scores across all six workload dimensions, indicating minimal mental ($M = 2.83$, $SD = 1.47$), physical ($M = 1.83$, $SD = 1.03$) and temporal ($M = 2.08$, $SD = 1.68$) demand; high performance ($M = 2.17$, $SD = 1.11$); and low effort ($M = 2.00$, $SD = 0.60$) and frustration ($M = 1.33$, $SD = 0.65$) (Figure 8). Mental demand, though still low, was the highest dimension. During the task-based usability test, 10 of 12 participants (P01, P02, P04, P05, P06, P08, P09, P10, P11, P12) described difficulty learning the system, with P06 commenting, "The hard thing was remembering," when asked about ease of use. All but one participant (P08) experienced

Participant ID	Medieval Market				Fast Food Restaurant			
	PO	CO	SC	SA	PO	CO	SC	SA
P01	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
P02	Pass	Pass	Pass	Fail	Pass	Pass	Pass	Pass
P03	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
P04	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
P05	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
P06	Pass	Pass	Pass	Fail	Pass	Pass	Pass	Pass
P07	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Fail
P08	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
P09	Pass	Pass	Fail	Pass	Pass	Pass	Pass	Fail
P10	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
P11	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
P12	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass

Table 2: Pass/fail results for each participant in the task-based usability study. Each of the four tasks (1) Parent Object (PO), (2) Child Object (CO), (3) Spatial Context (SC), and (4) Spatial Awareness (SA) was repeated in both the medieval market and the fast food restaurant test environments. Overall, 91 of 96 tasks were completed successfully (94.8%).

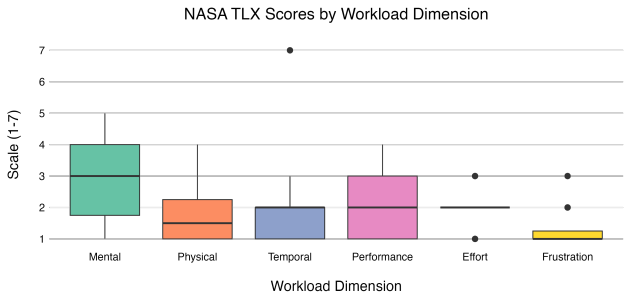


Figure 8: Distribution of NASA-TLX scores across each workload dimension. Each box represents the interquartile range (*IQR*), with the horizontal line inside the box indicating the median. Whiskers extend to $1.5 \times IQR$ beyond the first and third quartiles, and dots represent outliers, which were observed for temporal demand, effort and frustration. Lower scores indicate lower perceived workload when using SceneVR.

at least one moment where they forgot how to use the system and required assistance from the research team. However, eight participants (P01, P02, P04, P05, P06, P09, P11, P12) believed it would become easier with practice, as P01 remarked, "With practice ... it's easier ... [it] becomes something easy to do."

5.3 Sense of Presence

We used the iGroup Presence Questionnaire (*IPQ*) [70, 71] to measure participants' sense of presence in the virtual world when using SceneVR. The *IPQ* measures presence across three sub-scales: (1) spatial presence (SP), i.e., feeling physically present in the virtual world, (2) involvement (INV), i.e., attention focused on the virtual world, and (3) experienced realism (REAL), i.e., how lifelike the virtual environment feels. The *IPQ* also includes a general presence (GP) question measuring the overall "sense of being there." Most questions used a 1-7 Likert-type scale, with higher scores indicating a stronger perception of presence. For questions with an inverted scale, where 1 indicated the highest sense of presence, we reversed the scores for consistency during analysis.

Figure 9 shows the distribution of *IPQ* scores. Participants reported high general presence ($M = 5.50$, $SD = 1.73$) and spatial presence ($M = 5.35$, $SD = 1.66$). However, experienced realism was lower ($M = 3.33$, $SD = 1.88$), with some finding the virtual world less realistic, e.g., P02 noted that the environment still felt "virtual." Involvement scores were also moderate ($M = 4.10$, $SD = 2.20$), suggesting variation in the degree of engagement participants felt while interacting with the virtual scene.

5.4 Scene Reading Usage and Interaction Patterns

To better understand how participants engaged with scene reading and how touch-based interactions influenced their preferences, we analyzed participants' interaction patterns. First, we examined their use of scene reading methods by comparing the relative frequency of spatial (continuous finger-driven touch), sequential (discrete flicking), and overview (circle gesture) reading approaches. However, because spatial reading involves a continuous gesture while sequential and overview reading rely on discrete gestures, a direct count of interactions would not constitute a meaningful comparison.

To address this, we defined what constitutes a single instance of scene reading for each interaction type. For spatial reading, we defined an instance as beginning when the participant places their finger on the screen and ending when they lift it, with instances

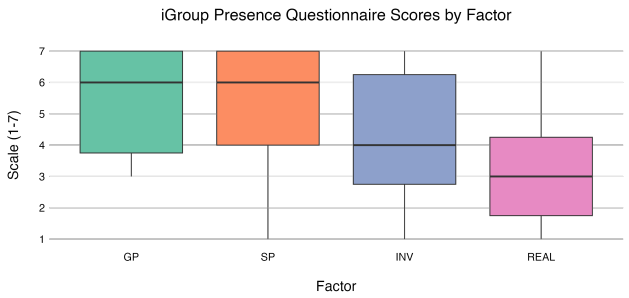


Figure 9: Distribution of *IPQ* scores across general presence (GP) and three sub-scales: spatial presence (SP), involvement (INV), and experienced realism (REAL). Each box represents the interquartile range (*IQR*), with the horizontal line inside the box indicating the median. Whiskers extend to $1.5 \times IQR$ beyond the first and third quartiles. No outliers were observed. Higher scores indicate a stronger sense of presence in the virtual world when using SceneVR.

under one second discarded as likely unintentional. For sequential reading, an instance starts with the first flick to move to another object and ends when the participant performs a different gesture or pauses for more than five seconds. This approach grouped rapid, successive flicks into a single scene reading instance, allowing for a more consistent comparison between continuous and discrete interaction methods. For overview reading, each activation of the one-finger circle gesture is considered a single instance.

Using this approach, we found that, on average, sequential reading accounted for 50.91% ($SD = 20.67$) of scene reading interactions per user, while overview reading accounted for 27.59% ($SD = 20.03$), and spatial reading accounted for 21.50% ($SD = 12.14$). This variation in interaction patterns was also reflected in qualitative feedback, with participants expressing different preferences for scene reading methods. Six of 12 participants (P03, P04, P06, P09, P10, P11) commented that the one-finger flick for sequential reading felt easy or was their preferred method of scene reading, while three participants (P02, P07, P12) expressed similar preferences for the one-finger circle used in overview reading. Although spatial scene reading accounted for an average of 21.50% of scene reading interactions, only one participant (P08) described it as their preferred method but noted that it became challenging when searching for objects that were farther away.

We also analyzed which types of annotation content participants accessed most frequently while scene reading. Expectedly, shorter object labels accounted for an average of 97.32% ($SD = 2.47$) of total annotation usage per participant, with longer object descriptions comprising the remaining proportion.

Figure 10 presents the mean relative frequency of these usage patterns, including scene reading interaction techniques (Figure 10a) and types of annotations accessed (Figure 10b).

5.5 Usability Challenges in a Multi-Sensory Environment

We observed that in a multi-sensory virtual environment, sensory cues can unexpectedly reveal gaps in scene reading annotation coverage. For instance, P06, who has some residual vision, saw a table with visible items but no accompanying annotations for those objects, prompting the question, "Is there stuff on the table though that I'd want to look at or there's nothing?" Similarly, P10 noticed an object in a certain direction without a label. Other expectations stemmed from contextual awareness, like those of P12, who, while exploring outdoors, wanted to access an annotation for the sky, asking, "How about the sky? Can I see [the annotation]?" In these cases, sensory input, such as visual details or environmental context, was followed by participants attempting to access annotations that were not available.

In other cases, participants commented on the absence or misalignment of expected sensory feedback following the use of scene-reading annotations. For example, P03 appreciated the system's spatial audio, particularly the verbal cues, but noted that the experience would be enhanced with richer ambient sounds, such as cues reflecting the time of day or natural sounds of birds in the virtual campground. Additionally, P02 reported that while the verbal feedback indicated a rightward turn, the continuous nonverbal spatial audio cue intended to evoke the feeling of turning failed to produce that sensation, resulting in an unrealistic and confusing orientation experience.

6 Discussion

Making highly visual and immersive 3-D virtual environments accessible to BLV people is a difficult design challenge, with few, if any, successful solutions developed to-date. Even major commercial manufacturers have not succeeded at this challenge. Our findings show that scene reading with SceneVR effectively enabled exploration and understanding of virtual scenes for BLV users. Participants successfully completed the task-based usability assessment, reported a strong sense of presence, and gave high satisfaction ratings. Qualitative feedback reinforced these results, with many describing the system as enjoyable and emphasizing their ability to use SceneVR independently, something they could not do with today's commercial stock controllers and software. This combination of enjoyment and sense of agency supports participants' emotional engagement with the technology [3]. Below, we unpack these insights by examining key findings from our study, highlighting what worked well and highlighting opportunities for further improvement.

6.1 Scene Reading and the Role of Object-Level Annotations

This work sought to address how we can reveal semantic information about objects and their on-screen position to help BLV users explore and understand virtual scenes. Our findings inform initial design implications and raise important questions for future research.

Sensory feedback and annotations are tightly coupled. When SceneVR users perceived an element, whether through limited vision, environmental context, or spatial audio, they often

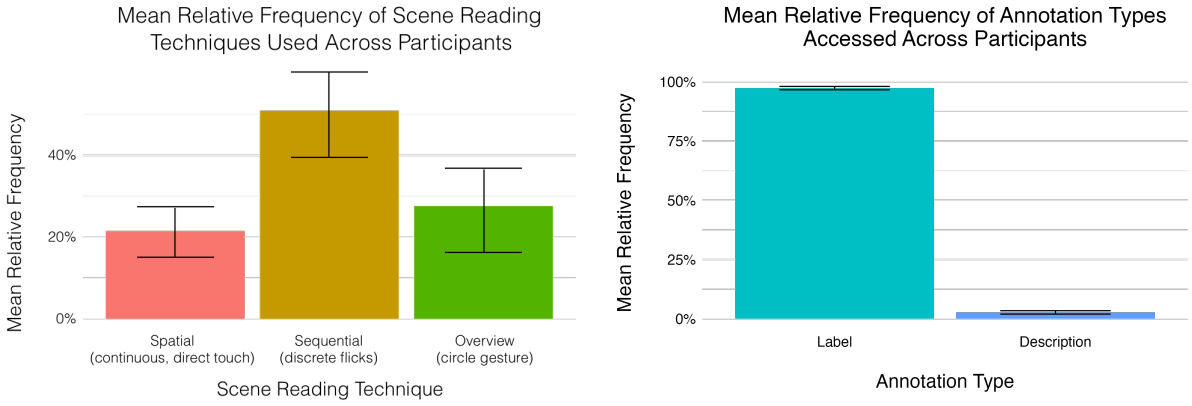


Figure 10: Mean relative frequency of scene reading interaction methods and types of object annotations accessed across participants in our SceneVR usability study. Error bars indicate $+/-1$ standard deviation (SD) for each mean. (a) When scene reading, participants most frequently used sequential reading ($M = 50.91\%, SD = 20.67$) to flick through objects in order, followed by overview (circle gesture) reading ($M = 27.59\%, SD = 20.03$) to identify all objects in view, and least often, spatial reading ($M = 21.50\%, SD = 12.14$) to explore objects by continuous, direct touch. (b) Short object labels accounted for the vast majority of annotations accessed during the usability study ($M = 97.32\%, SD = 2.47$), while longer object descriptions were used infrequently ($M = 2.68\%, SD = 2.47$).

expected a corresponding scene-reading annotation to confirm or clarify what they sensed. In some cases, such as P06 and P12, the details they became aware of had not been annotated and couldn't be accessed through SceneVR. Although our intent was to comprehensively annotate the scene, users' varied exploration strategies surfaced annotation gaps we had not anticipated. These gaps were not necessarily omissions of universally important details but rather reflected individual differences in what users perceived and wanted to explore further (e.g., the sky). As prior work in multi-sensory VR has emphasized, users do not share a single, uniform representation of a scene, and how scenes are perceived can vary widely across individuals [7]. This insight suggests that designers cannot reliably predict which elements will shape users' understanding of a virtual environment and underscores the need for comprehensive object annotations to support sensory-driven exploration.

Conversely, annotations themselves can shape expectations for additional sensory feedback. When users accessed an annotation for an element typically associated with environmental cues, they often expected those sensory details to be present in the environment. P02 and P03, for example, noticed when such cues were missing and suggested adding ambient spatial audio to reinforce the accessibility information provided by SceneVR. These moments suggest a disconnect between what is described and what is experienced when expected sensory feedback is absent or misaligned with scene-reading annotations, which may have contributed to lower *IPQ* scores for realism and involvement.

To create a more cohesive experience, multi-sensory virtual environments should ensure that scene-reading annotations and sensory feedback work together: any element users might perceive should be annotated, and any annotated element should include sensory feedback that deepens the experience. Rather than functioning independently, annotations should complement other system cues, such as environmental audio or haptic feedback, which also influence how users perceive and make sense of a scene. Supporting this interplay can better meet BLV user expectations and enable exploration grounded in individual sensory experiences.

Object labels are a key source of information during virtual exploration. Participants in our study consistently relied upon short-form object labels, highlighting their importance in scene exploration and understanding. By contrast, long-form object descriptions were used much less frequently. Since accessing descriptions required an additional gesture (the split-tap), lower usage is unsurprising. Additionally, their lower usage may reflect the nature of our assigned tasks, which did not emphasize the need for detailed descriptions. Given these uncertainties, we refrain from drawing firm conclusions about the role of long-form object descriptions. Instead, our findings emphasize that short-form labels play a critical role in providing immediate, object-level information. To support efficient exploration and understanding, scene-reading techniques should prioritize making these labels easily and quickly accessible.

6.2 Touch-Based Techniques, Progressive Disclosure, and Managing Cognitive Load

This work also sought to address how to adapt touch-based interaction techniques and progressive disclosure of hierarchical information, designed for 2-D visual exploration, to support BLV understanding of 3-D virtual scenes. To approach this challenge, we examine the role that these techniques play in managing cognitive load and facilitating effective exploration. While these techniques

may help structure exploration and information access, they also introduce additional complexity in 3-D environments.

Structured exploration methods complement free-form spatial exploration. Participants engaged with the environment using a mix of spatial (continuous, direct touch), sequential (discrete flicks), and overview (circle gesture) scene reading. Structured methods (sequential and overview reading, which provide access to objects in a predefined and predictable order) were used more frequently and were more commonly described as easier to use or preferred. This finding aligns with prior research in 2-D image exploration [57] and BLV video game [55] and AR [33] accessibility, which found that structured, menu-based navigation was favored for ease of use and reliably accessing all objects, while free-form exploration provided greater autonomy and spatial awareness, but was more difficult to control.

Although spatial scene reading was used less frequently overall, it still accounted for an average of 21.50% ($SD = 12.14$) of scene reading interactions, indicating its value to participants. We hypothesize a trade-off similar to that found in prior work: structured techniques offer predictability and ease of access, while spatial techniques support autonomy and spatial understanding. To accommodate diverse user needs, scene-reading systems should integrate multiple exploration methods, balancing predictability, efficiency, and exploratory freedom.

Progressive disclosure of detail through hierarchical object organization helps manage information overload. Virtual environments contain many objects, often with dynamic interactions and complex relationships, making nonvisual exploration cognitively demanding. Without a way to structure this information, users may have to process many details at once, increasing cognitive load and complicating efforts to find relevant objects. Progressive disclosure of increasing level-of-detail (LOD) through proximity-based and object-group hierarchy navigation helps mitigate information overload by limiting the number of objects presented at one time and supporting an intuitive search strategy where users first locate a parent object and then navigate its hierarchy to find related items.

To wit, all participants successfully completed the child object (CO) task in our usability study, suggesting that our progressive disclosure techniques did not interfere with object discovery and were usable in practice. Additionally, participants specifically highlighted object-group disclosure as particularly useful for exploration. One possible reason is that they valued the enforced focus provided by object groups. By automatically centering the scene-reading view on relevant objects within a group, object groups reduce the burden of manual positioning for efficient scene reading. Together, these findings indicate that structuring scene reading through hierarchical navigation, with built-in support for scene-reading focus, can reduce cognitive load and improve access to fine scene detail.

Touch-based interaction is essential for spatial scene reading, but may present challenges when scaling to more complex virtual environments. Spatial scene reading relies on direct touch input, and participants' use of this technique throughout the study reinforces its role as a core component of the broader scene reading toolkit. However, while participants successfully engaged with scene reading overall, many struggled to remember the full set of touch gestures, with nearly all participants experiencing at least one moment where they forgot how to use a desired gesture. Although several participants speculated that the system would become easier with practice, their initial difficulty also raises concerns about cognitive load, particularly during early use. These findings prompt questions about the scalability of similar touch-based techniques in more complex VR environments that demand additional interactions, such as object manipulation or social engagement.

7 Limitations and Future Work

Our study examined how scene reading, supported by object hierarchies, progressive disclosure, and touch-based techniques, contributed to BLV users' experiences of virtual environments. Naturally, this work required designing a specific interaction technique for accessing object annotations. Our approach was informed by prior work, but we recognize that testing a single design does not necessarily determine the most effective approach to scene reading in general.

Originally, we planned to include a comparison condition that would have enabled spatial scene reading using a ray casting technique with a commercial stock VR controller. However, our pilot testing revealed that naively enabling VR controllers to read object labels via ray casting did not work well. Because commercial VR headsets lack robust accessibility tools for BLV users, there was no *de facto* baseline for participants to rely on, and everything in the study, from object labeling to exploration methods, had to be learned from scratch. As a result, participants required more time to learn a single system, and introducing SceneVR alongside additionally novel comparison conditions proved too complex for a single study. Although a comparative study remains important for future work, as does making stock VR controllers accessible with commercial VR environments, we chose to begin with an original evaluation of scene reading and touch-based interactions in SceneVR. This approach allowed us to explore the feasibility of our approach, examine how users engaged with features, and surface initial design considerations that can inform and guide future research.

Future work should indeed conduct a formal comparison of additional interaction techniques and alternative gestures, particularly those used for spatial scene reading. For example, future research could compare direct touch input, ray casting, and in-air gestures to assess their relative effectiveness for object discovery and selection, for navigation, and for fostering spatial awareness.

In parallel, more work is needed to understand how hierarchical structures scale in practice. While our study explored object hierarchies and progressive disclosure of detail, future work should take a closer look at how these approaches scale. For example, how many levels of hierarchy remain usable before the structure becomes confusing, and how can we determine whether objects are grouped into intuitive, "natural" hierarchies? The scenes used in our study featured relatively straightforward object structures, but more complex or unfamiliar environments may present new challenges that warrant further investigation.

Effectively managing object hierarchies at scale will also require robust methods for generating and maintaining object annotations. Although object annotation was out of scope for this work, future research should evaluate whether the object attributes we used are

sufficient to support accessibility across larger and more complex environments. In addition, work is needed to explore how these attributes could be efficiently supplied at scale, potentially using automated techniques, such as large language models (LLMs), to generate object labels, descriptions, and intuitive hierarchies that reflect how users naturally explore a scene. Future systems will also need to support flexible access methods, such as structured and spatial exploration techniques; future work should also better understand the complexity these demands add to underlying VR infrastructure and interaction design requirements.

Finally, future work should examine the role and authorship of object descriptions. In our study, participants accessed object labels far more frequently than object descriptions, but the reasons for this behavior remain unclear. Further research should investigate when and why object descriptions are valuable, and how best to write them for 3-D virtual environments.

Our research provides an initial foundation for understanding scene reading and how to enable it through touch-based interaction techniques. However, since these techniques are relatively new, future research should explore these and other aspects in more detail to optimize both information and interaction design.

8 Conclusion

In this work, we have introduced "scene reading" as an analog to screen reading but for 3-D virtual environments, not 2-D web pages and user interfaces. Our scene reading interaction techniques, which rely on touch, gesture, and spatial audio, were explored in our *SceneVR* prototype, which we evaluated with 12 BLV participants in a task-based usability test. Our primary goal was to enable BLV users to explore and understand virtual scenes through nonvisual access to semantic information about objects and their on-screen positions. This goal was achieved, as participants were 94.8% successful at completing tasks in our virtual environments using SceneVR.

SceneVR was a touchscreen VR controller running on a horizontally oriented Apple iOS phone or tablet device, coupled with a Meta Quest 2 VR headset. To facilitate efficient scene and object exploration, scene reading in SceneVR relied upon object hierarchies that users could navigate to progressively discover more detail. Along with successful task completion, our participants reported SceneVR providing a strong sense of presence and enhancing user enjoyment and agency. Participants felt they could operate SceneVR independently, which was not the case for current commercial VR controllers or environments.

Beyond these benefits, our findings highlight initial design considerations. Feedback in multi-sensory environments shapes users' expectations for annotations and vice versa, emphasizing the need for comprehensive annotations that align with and complement sensory cues. Our findings also reveal that progressive disclosure through hierarchical object organization and navigation helps manage information overload.

Overall, our research demonstrates the promise of touch-based hierarchical scene reading techniques to enable BLV exploration of virtual environments while also offering a clearer understanding of the challenges and future research necessary to refine and scale these techniques for broader applicability.

Acknowledgments

The authors thank Arnavi Chheda-Kothary and Sandy Kaplan. This work was supported in part by an award from Facebook on Social Experiences in VR Environments and by the University of Washington Center for Research and Education on Accessible Technology and Experiences (CREATE). Any opinions, findings, conclusions or recommendations expressed in our work are those of the authors and do not necessarily reflect those of any supporter. ChatGPT was utilized to generate aspects of this work, including text, tables, and graphs, but was not used to generate original text or ideas, conduct data analyses, or make findings, which are solely the work of the authors.

References

[1] Dragan Ahmetovic, Nahyun Kwon, Uran Oh, Cristian Bernareggi, and Sergio Mascetti. 2021. Touch Screen Exploration of Visual Artwork for Blind People. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 2781–2791. doi:10.1145/3442381.3449871

[2] Denys Almaral. 2024. City People FREE Samples. https://assetstore.unity.com/packages/3d/characters/city-people-free-samples-260446.

[3] Oqab Alrashidi, Huy P. Phan, and Bing H. Ngu. 2016. Academic Engagement: An Overview of Its Definitions, Dimensions, and Major Conceptualisations. *International Education Studies* 9, 12 (Nov. 2016), 41. doi:10.5539/ies.v9n12p41

[4] American Foundation for the Blind. 2024. Screen readers.

[5] Ronny Andrade, Steven Baker, Jenny Waycott, and Frank Vetere. 2018. Echohouse: exploring a virtual environment by using echolocation. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*. ACM, Melbourne Australia, 278–289. doi:10.1145/3292147.3292163

[6] Teo Babic, Harald Reiterer, and Michael Haller. 2018. Pocket6: A 6DoF Controller Based On A Simple Smartphone Application. In *Proceedings of the Symposium on Spatial User Interaction*. ACM, Berlin Germany, 2–10. doi:10.1145/3267782.3267785

[7] Harshadha Balasubramanian, Cecily Morrison, Martin Grayson, Zhanat Makhataeva, Rita Faia Marques, Thomas Gable, Dalya Perez, and Edward Cutrell. 2023. Enable Blind Users' Experience in 3D Virtual Environments: The Scene Weaver Prototype. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–4. doi:10.1145/3544549.3583909

[8] Rhyse Bendell and Jessica Williams. 2023. Assessing Spatial Knowledge and Mental Map Development Under Virtual Training Conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67, 1 (Sept. 2023), 1611–1616. doi:10.1177/21695067231192219

[9] Verena Biener, Daniel Schneider, Travis Gesslein, Alexander Otte, Bastian Kuth, Per Ola Kristensson, Eyal Ofek, Michel Pahud, and Jens Grubert. 2020. Breaking the Screen: Interaction Across Touchscreen Boundaries in Virtual Reality for Mobile Knowledge Workers. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (Dec. 2020), 3490–3502. doi:10.1109/TVCG.2020.3023567

[10] Blink. 2022. FREE Stylized Bear - RPG Forest Animal. https://assetstore.unity.com/packages/3d/characters/animals/free-stylized-bear-rpg-forest-animal-228910.

[11] Sabah Boustila, Thomas Guegan, Kazuki Takashima, and Yoshifumi Kitamura. 2019. Text Typing in VR Using Smartphones Touchscreen and HMD. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Osaka, Japan, 860–861. doi:10.1109/VR.2019.8798238

[12] Doug A. Bowman, Joseph L. Gabbard, and Deborah Hix. 2002. A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. *Presence: Teleoperators and Virtual Environments* 11, 4 (Aug. 2002), 404–424. doi:10.1162/105474602760204309

[13] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[15] William Carter and Guido Corona. 2008. Exploring Methods of Accessing Virtual Worlds. *AccessWorld* 9, 2 (2008). https://www.afb.org/aw/9/2/14262

[16] Junlong Chen, Rosella P. Galindo Esparza, Vanja Garaj, Per Ola Kristensson, and John Dudley. 2025. EnVisionVR: A Scene Interpretation Tool for Visual Accessibility in Virtual Reality. doi:10.48550/ARXIV.2502.03564 Version Number: 1.

[17] Jazmin Collins, Crescentia Jung, and Shiri Azenkot. 2023. Making Avatar Gaze Accessible for Blind and Low Vision People in Virtual Reality: Preliminary Insights. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, Sydney, Australia, 701–705. doi:10.1109/ISMAR-

Adjunct60411.2023.00150

[18] Chetz Colwell, Helen Petrie, Diana Kornbrot, Andrew Hardwick, and Stephen Furner. 1998. Haptic virtual reality for blind computer users. In *Proceedings of the third international ACM conference on Assistive technologies*. ACM, Marina del Rey California USA, 92–99. doi:10.1145/274497.274515

[19] Maurizio De Pascale, Sara Mulatto, and Domenico Prattichizzo. 2008. Bringing Haptics to Second Life for Visually Impaired People. In *Haptics: Perception, Devices and Scenarios*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, and Manuel Ferre (Eds.). Vol. 5024. Springer Berlin Heidelberg, Berlin, Heidelberg, 896–905. doi:10.1007/978-3-540-69057-3_112 Series Title: Lecture Notes in Computer Science.

[20] Equal Entry. 2022. Virtual Reality Accessibility: 11 Things We Learned from Blind Users. https://equalentry.com/virtual-reality-accessibility-things-learned-from-blind-users/.

[21] Steven Feiner, Blair Maclntyre, Tobias Höllerer, and Anthony Webster. 1997. A Touring Machine: Prototyping 3D Mobile Augmented Reality Systems for Exploring the Urban Environment. *Personal Technologies* 1 (1997), 208–217.

[22] Eelke Folmer, Bei Yuan, Dave Carr, and Manjari Sapre. 2009. TextSL: a command-based virtual world interface for the visually impaired. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. ACM, Pittsburgh Pennsylvania USA, 59–66. doi:10.1145/1639642.1639654

[23] Rachel L. Franz, Sasa Junuzovic, and Martez Mott. 2024. A Virtual Reality Scene Taxonomy: Identifying and Designing Accessible Scene-Viewing Techniques. *ACM Transactions on Computer-Human Interaction* 31, 2 (April 2024), 1–44. doi:10.1145/3635142

[24] Aaron Gluck, Kwajo Boateng, and Julian Brinkley. 2021. Racing in the Dark: Exploring Accessible Virtual Reality by Developing a Racing Game for People who are Blind. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65, 1 (Sept. 2021), 1114–1118. doi:10.1177/1071181321651224

[25] Cagatay Goncu and Kim Marriott. 2011. GraVVITAS: Generic Multi-touch Presentation of Accessible Graphics. In *Human-Computer Interaction – INTERACT 2011*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 30–48.

[26] Ricardo E. Gonzalez Penuela, Wren Poremba, Christina Trice, and Shiri Azenkot. 2022. Hands-On: Using Gestures to Control Descriptions of a Virtual Environment for People with Visual Impairments. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, Bend OR USA, 1–4. doi:10.1145/3526114.3558669

[27] Jens Grubert, Matthias Heinisch, Aaron Quigley, and Dieter Schmalstieg. 2015. MultiFi: Multi Fidelity Interaction with Displays On and Around the Body. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 3933–3942. doi:10.1145/2702123.2702331

[28] João Guerreiro, Yujin Kim, Rodrigo Nogueira, SeungA Chung, André Rodrigues, and Uran Oh. 2023. The Design Space of the Auditory Representation of Objects and Their Behaviours in Virtual Reality for Blind People. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (May 2023), 2763–2773. doi:10.1109/TVCG.2023.3247094

[29] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. doi:10.1177/154193120605000909

[30] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183. doi:10.1016/S0166-4115(08)62386-9

[31] Harvard University. 2025. Write helpful Alt Text to describe images. https://accessibility.huit.harvard.edu/describe-content-images.

[32] Alex Heath. 2021. Meta opens up access to its VR social platform Horizon Worlds. https://www.theverge.com/2021/12/9/22825139/meta-horizon-worlds-access-open-metaverse. *The Verge* (09 Dec. 2021).

[33] Jaylin Herskovitz, Jason Wu, Samuel White, Amy Pavel, Gabriel Reyes, Anhong Guo, and Jeffrey P. Bigham. 2020. Making Mobile Augmented Reality Applications Accessible. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Virtual Event Greece, 1–14. doi:10.1145/3373625.3417006

[34] Sebastian Hubenschmid, Johannes Zagermann, Simon Butscher, and Harald Reiterer. 2021. STREAM: Exploring the Combination of Spatially-Aware Tablets with Augmented Reality Head-Mounted Displays for Immersive Analytics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3411764.3445298

[35] Tero Jokela and Andrés Lucero. 2014. MixedNotes: a digital tool to prepare physical notes for affinity diagramming. In *Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services*. ACM, Tampere Finland, 3–6. doi:10.1145/2676467.2676478

[36] Ankur Joshi, Saket Kale, Satish Chandel, and D. Pal. 2015. Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology* 7, 4 (Jan. 2015), 396–403. doi:10.9734/BJAST/2015/14975

[37] Crescentia Jung, Jazmin Collins, Ricardo E. Gonzalez Penuela, Jonathan Isaac Segal, Andrea Stevenson Won, and Shiri Azenkot. 2024. Accessible Nonverbal Cues to Support Conversations in VR for Blind and Low Vision People. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, St. John's NL Canada, 1–13. doi:10.1145/3663548.3675663

[38] Shaun K. Kane, Jeffrey P. Bigham, and Jacob O. Wobbrock. 2008. Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. ACM, Halifax Nova Scotia Canada, 73–80. doi:10.1145/1414471.1414487

[39] Shaun K. Kane, Brian Frey, and Jacob O. Wobbrock. 2013. Access lens: a gesture-based screen reader for real-world documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 347–350. doi:10.1145/2470654.2470704

[40] Shaun K. Kane, Meredith Ringel Morris, Annuska Z. Perkins, Daniel Wigdor, Richard E. Ladner, and Jacob O. Wobbrock. 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*. ACM, Santa Barbara California USA, 273–282. doi:10.1145/2047196.2047232

[41] Mohamed Kari and Christian Holz. 2023. HandyCast: Phone-based Bimanual Input for Virtual Reality in Mobile and Space-Constrained Settings via Pose-and-Touch Transfer. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. doi:10.1145/3544548.3580677

[42] Diana Kornbrot, Paul Penn, Helen Petrie, Stephen Furner, and Andrew Hardwick. 2007. Roughness perception in haptic virtual reality for sighted and blind people. *Perception & Psychophysics* 69, 4 (May 2007), 502–512. doi:10.3758/BF03193907

[43] Rynhardt Kruger and Lynette van Zijl. 2015. Virtual World Accessibility with the Perspective Viewer. In *ICEAPVI*. Athens, Greece.

[44] Ricardo Langner, Marc Satkowski, Wolfgang Büschel, and Raimund Dachselt. 2021. MARVIS: Combining Mobile Devices and Augmented Reality for Visual Data Analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–17. doi:10.1145/3411764.3445593

[45] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. doi:10.1145/3491102.3501966

[46] Jaewook Lee, Yi-Hao Peng, Jaylin Herskovitz, and Anhong Guo. 2021. Image Explorer: Multi-Layered Touch Exploration to Make Images Accessible. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Virtual Event USA, 1–4. doi:10.1145/3441852.3476548

[47] Joon Hyub Lee, Taegyu Jin, Sang-Hyun Lee, Seung-Jun Lee, and Seok-Hyung Bae. 2023. Stereoscopic Viewing and Monoscopic Touching: Selecting Distant Objects in VR Through a Mobile Device. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–7. doi:10.1145/3586183.3606809

[48] Anatole Lécuyer, Pascal Mobuchon, Christine Mégard, Jérôme Perret, Claude Andriot, and Jean-Pierre Colinot. 2003. HOMERE: a multimodal system for visually impaired people to explore virtual environments. (March 2003), 251–258. doi:10.1109/vr.2003.1191147 MAG ID: 2122069282.

[49] Bernard Marr. 2023. Game On! The Top 10 Video Game Trends In 2024. https://www.forbes.com/sites/bernardmarr/2023/09/29/game-on-the-top-10-video-game-trends-in-2024/?sh=1518f80e381d. *Forbes* (29 Sept. 2023).

[50] Fabrice Matulic, Aditya Ganeshan, Hiroshi Fujiwara, and Daniel Vogel. 2021. Phonetroller: Visual Representations of Fingers for Precise Touch Input with Mobile Phones in VR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3411764.3445583

[51] Peter Mohr, Markus Tatzgern, Tobias Langlotz, Andreas Lang, Dieter Schmalstieg, and Denis Kalkofen. 2019. TrackCap: Enabling Smartphones for 3D Interaction on Mobile Head-Mounted Displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. doi:10.1145/3290605.3300815

[52] Tony Morelli, John Foley, Luis Columna, Lauren Lieberman, and Eelke Folmer. 2010. VI-Tennis: a vibrotactile/audio exergame for players who are visually impaired. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*. ACM, Monterey California, 147–154. doi:10.1145/1822348.1822368

[53] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–11. doi:10.1145/3173574.3173633

[54] Martez Mott, Edward Cutrell, Mar Gonzalez Franco, Christian Holz, Eyal Ofek, Richard Stoakley, and Meredith Ringel Morris. 2019. Accessible by Design: An Opportunity for Virtual Reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, Beijing, China, 451–454. doi:10.1109/ISMAR-Adjunct.2019.00122

[55] Vishnu Nair, Jay L Karp, Samuel Silverman, Mohar Kalra, Hollis Lehv, Faizan Jamil, and Brian A. Smith. 2021. NavStick: Making Video Games Blind-Accessible via the Ability to Look Around. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 538–551. doi:10.1145/3472749.3474768

[56] Vishnu Nair, Shao-en Ma, Ricardo E. Gonzalez Penuela, Yicheng He, Karen Lin, Mason Hayes, Hannah Huddleston, Matthew Donnelly, and Brian A. Smith. 2022. Uncovering Visually Impaired Gamers' Preferences for Spatial Awareness Tools Within Video Games. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Athens Greece, 1–16. doi:10.1145/3517428.3544802

[57] Vishnu Nair, Hanxiu 'Hazel' Zhu, and Brian A. Smith. 2023. ImageAssist: Tools for Enhancing Touchscreen-Based Image Exploration Systems for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. doi:10.1145/3544548.3581302

[58] Vishnu Nair, Hanxiu 'Hazel' Zhu, Peize Song, Jizhong Wang, and Brian A. Smith. 2024. Surveyor: Facilitating Discovery Within Video Games for Blind and Low Vision Players. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3613904.3642615

[59] NeutronCat. 2020. Simple Low Poly Nature Pack. https://assetstore.unity.com/packages/3d/environments/landscapes/simple-low-poly-nature-pack-157552.

[60] Jakob Nielsen. 1993. *Usability engineering*. Academic Press, Boston.

[61] Jakob Nielsen. 2006. Progressive Disclosure. https://www.nngroup.com/articles/progressive-disclosure/.

[62] Georgios Nikolakis, Dimitrios Tzovaras, Serafim Moustakidis, and Michael G Strintzis. 2004. CyberGrasp and PHANTOM Integration: Enhanced Haptic Access for Visually Impaired Users. In *Proceedings of the 9th Conference on Speech and Computer*. St. Petersburg, Russia, 1–7.

[63] Yuki Noguchi. 2019. Virtual Reality Goes To Work, Helping Train Employees. https://www.npr.org/2019/10/08/767116408/virtual-reality-goes-to-work-helping-train-employees. *NPR* (08 Oct. 2019).

[64] Chris Nolet. 2022. Quick Outline. https://assetstore.unity.com/packages/tools/particles-effects/quick-outline-115488.

[65] Bugra Oktay and Eelke Folmer. 2010. Synthesizing meaningful feedback for exploring virtual worlds using a screen reader. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 4165–4170. doi:10.1145/1753846.1754120

[66] Bugra Oktay and Eelke Folmer. 2011. Syntherella: a feedback synthesizer for efficient exploration of virtual worlds using a screen reader. In *Proceedings of Graphics Interface 2011 (GI 2011)*. Canadian Human-Computer Communications Society, St. John's, Newfoundland, Canada, 65–70.

[67] David Redmond. 2024. Where's the screen reader? How the Meta Quest Pro could be made more accessible. https://vi.ie/wheres-the-screen-reader-how-the-meta-quest-pro-could-be-made-more-accessible/. *Vision Ireland* (03 May 2024).

[68] Sol Rogers. 2020. How Virtual Reality Could Help The Travel & Tourism Industry In The Aftermath Of The Coronavirus Outbreak. https://www.forbes.com/sites/solrogers/2020/03/18/virtual-reality-and-tourism-whats-already-happening-is-it-the-future/. *Forbes* (18 March 2020).

[69] Anastasia Schaadhardt, Alexis Hiniker, and Jacob O. Wobbrock. 2021. Understanding Blind Screen-Reader Users' Experiences of Digital Artboards. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3411764.3445242

[70] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments* 10, 3 (06 2001), 266–281. arXiv:https://direct.mit.edu/pvar/article-pdf/10/3/266/1623697/105474601300343603.pdf doi:10.1162/105474601300343603

[71] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2025. iGroup Presence Questionnaire (IPQ). https://www.igroup.org/pq/ipq.

[72] Ather Sharif, Venkatesh Potluri, Jazz Rui Xia Ang, Jacob O. Wobbrock, and Jennifer Mankoff. 2024. Touchpad Mapper: Examining Information Consumption From 2D Digital Content Using Touchpads by Screen-Reader Users. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, St. John's NL Canada, 1–4. doi:10.1145/3663548.3688505

[73] Ather Sharif, Andrew M. Zhang, Katharina Reinecke, and Jacob O. Wobbrock. 2023. Understanding and Improving Drilled-Down Information Extraction from Online Data Visualizations for Screen-Reader Users. In *20th International Web for All Conference*. ACM, Austin TX USA, 18–31. doi:10.1145/3587281.3587284

[74] Alexa F. Siu, Mike Sinclair, Robert Kovacs, Eyal Ofek, Christian Holz, and Edward Cutrell. 2020. Virtual Reality Without Vision: A Haptic and Auditory White Cane to Navigate Complex Virtual Worlds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376353

[75] Sharon Spall. 1998. Peer Debriefing in Qualitative Research: Emerging Operational Models. *Qualitative Inquiry* 4, 2 (June 1998), 280–292. doi:10.1177/107780049800400208

[76] Stasel. 2018. Open Source WebRTC for iOS. https://github.com/stasel/WebRTC-iOS.

[77] Brick Project Studio. 2023. Fast Food Restaurant Kit. https://assetstore.unity.com/packages/3d/environments/fast-food-restaurant-kit-239419.

[78] Winged Boots Studio. 2023. Stylized NPC - Peasant Nolant (DEMO). https://assetstore.unity.com/packages/3d/characters/humanoids/fantasy/stylized-npc-peasant-nolant-demo-252440.

[79] Yu Sun, Carolin Stellmacher, Annika Kaltenhauser, Nadine Wagener, Daniel Neumann, and Johannes Schöning. 2023. Alt Text and Alt Sense in VR: Engaging Screen Reader Users within the Metaverse Through Multisenses. (2023).

[80] Zsolt Szalavári and Michael Gervautz. 1997. The Personal Interaction Panel – a Two-Handed Interface for Augmented Reality. *Eurographics* 16, 3 (1997), C335–C346. doi:10.1111/1467-8659.00137

[81] Mai Ricaplaza Thøgersen and Rasmus Jens Frølich Kjeldsen. 2024. Echolocation as an Accessible Navigation Tool in a Virtual 3D Environment. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, St. John's NL Canada, 1–9. doi:10.1145/3663548.3688547

[82] Miguel Torres Gil, Oscar Casanova, and Jose González-Mora. 2010. Applications of Virtual Reality for Visually Impaired People. *WSEAS Transactions on Computers* 9 (Feb. 2010).

[83] Shari Trewin, Vicki L. Hanson, Mark R. Laff, and Anna Cavender. 2008. PowerUp: an accessible virtual world. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. ACM, Halifax Nova Scotia Canada, 177–184. doi:10.1145/1414471.1414504

[84] Shari Trewin, Mark Laff, Vicki Hanson, and Anna Cavender. 2009. Exploring Visual and Motor Accessibility in Navigating a Virtual World. *ACM Transactions on Accessible Computing* 2, 2 (June 2009), 1–35. doi:10.1145/1530064.1530069

[85] Dimitrios Tzovaras, Konstantinos Moustakas, Georgios Nikolakis, and Michael G. Strintzis. 2009. Interactive mixed reality white cane simulation for the training of the blind and the visually impaired. *Personal and Ubiquitous Computing* 13, 1 (Jan. 2009), 51–58. doi:10.1007/s00779-007-0171-2

[86] Dimitrios Tzovaras, Georgios Nikolakis, Georgios Fergadis, Stratos Malasiotis, and Modestos Stavrakis. 2004. Design and implementation of haptic virtual environments for the training of the visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 12, 2 (June 2004), 266–278. doi:10.1109/TNSRE.2004.828756

[87] Arda Ege Unlu and Robert Xiao. 2021. PAIR: Phone as an Augmented Immersive Reality Controller. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. ACM, Osaka Japan, 1–6. doi:10.1145/3489849.3489878

[88] Gregg C. Vanderheiden. 1996. Use of Audio-Haptic Interface Techniques to Allow Nonvisual Access to Touchscreen Appliances. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, 24 (Oct. 1996), 1266–1266. doi:10.1177/154193129604002430

[89] VanillaArt. 2023. Low-Poly Medieval Market. https://assetstore.unity.com/packages/3d/environments/low-poly-medieval-market-262473.

[90] Shoshanah Wall. 2023. What is LOD (Level of Detail) in 3D Modeling? https://www.cgspectrum.com/blog/what-is-level-of-detail-lod-3d-modeling.

[91] Ryan Wedoff, Lindsay Ball, Amelia Wang, Yi Xuan Khoo, Lauren Lieberman, and Kyle Rector. 2019. Virtual Showdown: An Accessible Virtual Reality Game with Scaffolds for Youth with Visual Impairments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. doi:10.1145/3290605.3300371

[92] Jacob O Wobbrock, Rachel L Franz, and Melanie Kneitmix. 2024. Improving the accessibility of virtual reality for people with motor and visual impairments. In *Workshop on "Building a Metaverse for All: Opportunities and Challenges for Future Inclusive and Accessible Virtual Environments (Metaverse4All '24)"*. ACM, Honolulu, Hawaii, Paper No. 4.

[93] Zhuohao Zhang, John R Thompson, Aditi Shah, Manish Agrawal, Alper Sarikaya, Jacob O. Wobbrock, Edward Cutrell, and Bongshin Lee. 2024. ChartA11y: Designing Accessible Touch Experiences of Visualizations with Blind Smartphone Users. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, St. John's NL Canada, 1–15. doi:10.1145/3663548.3675611

[94] Zhuohao Zhang and Jacob O. Wobbrock. 2022. A11yBoard: Using Multimodal Input and Output to Make Digital Artboards Accessible to Blind Users. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, Bend OR USA, 1–4. doi:10.1145/3526114.3558695

[95] Zhuohao (Jerry) Zhang and Jacob O. Wobbrock. 2023. A11yBoard: Making Digital Artboards Accessible to Blind and Low-Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. doi:10.1145/3544548.3580655

[96] Yuhang Zhao, Cynthia L. Bennett, Hrvoje Benko, Edward Cutrell, Christian Holz, Meredith Ringel Morris, and Mike Sinclair. 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*. ACM, Montreal QC Canada, 1–14. doi:10.1145/3173574.3173690

[97] Yuhang Zhao, Edward Cutrell, Christian Holz, Meredith Ringel Morris, Eyal Ofek, and Andrew D. Wilson. 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*. ACM, Glasgow Scotland Uk, 1–14. doi:10.1145/3290605.3300341

[98] Fengyuan Zhu and Tovi Grossman. 2020. BISHARE: Exploring Bidirectional Interactions Between Smartphones and Head-Mounted Augmented Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. doi:10.1145/3313831.3376233