StreetViewAI: Making Street View Accessible Using Context-Aware Multimodal AI

Jon E. Froehlich
Google Research
jfroehlich@google.com

Alex Fiannaca
Google DeepMind
afiannaca@google.com

Nimer Jaber
Google
nimer@google.com

Victor Tsaran
Google
vtsaran@google.com

Shaun Kane
Google Research
shaunkane@google.com

Figure 1: We introduce StreetViewAI, an accessible streetscape mapping tool that uses context-aware AI and accessible navigation controls for blind and low-vision users. (a) Users can search for and select locations, (b) trigger AI-based descriptions, or (c) chat with a multimodal AI agent about the scene and local geography while virtually exploring the world.

Abstract

Interactive streetscape mapping tools such as Google Street View (GSV) and Meta Mapillary enable users to virtually navigate and experience real-world environments via immersive 360° imagery but remain fundamentally inaccessible to blind users. We introduce *StreetViewAI*, the first-ever accessible street view tool, which combines context-aware, multimodal AI, accessible navigation controls, and conversational speech. With StreetViewAI, blind users can virtually examine destinations, engage in open-world exploration, or virtually tour any of the over 220 billion images and 100+ countries where GSV is deployed. We iteratively designed StreetViewAI with a mixed-visual ability team and performed an evaluation with eleven blind users. Our findings demonstrate the value of an accessible street view in supporting POI investigations and remote route planning. We close by enumerating key guidelines for future work.

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747756

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI); Accessibility systems and tools**.

Keywords

Accessible maps, Multimodal LLMs, AI Chat, Street View

ACM Reference Format:

Jon E. Froehlich, Alex Fiannaca, Nimer Jaber, Victor Tsaran, and Shaun Kane. 2025. StreetViewAI: Making Street View Accessible Using Context-Aware Multimodal AI. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3746059.3747756

1 Introduction

Interactive, digital maps have transformed how people move about the world, from route planning and turn-by-turn directions to virtual tourism. Though historically inaccessible to blind users [34], recent work has significantly improved traditional two-dimensional

Figure 2: StreetViewAI's context-aware multimodal *AI Chat*. (a) The text prompt, optional user profile, and command function definitions are sent on "MLLM live chat" initiation. (b) The MLLM Live module and (c) panning/movement control system. (d) Once a chat is initiated, StreetViewAI uploads the user's current view and geographic context (*e.g.,* <lat,lng>, current heading, nearby places)—seamlessly updated every view change or movement. (e, f) Example context-aware AI chats. (g) The custom GSV-based navigational network. In this example, the user is virtually exploring a walk from a transit station to a cafe in NYC.

map accessibility through audio descriptions and spatialized audio [41, 90, 93], tactile representations [26, 65, 67, 96], and dynamic haptic feedback [31, 100, 132]. Despite this progress, an entire class of digital maps remains inaccessible: immersive streetscape mapping tools [9, 50, 89, 91]. These interactive tools enable users to virtually navigate and experience real-world environments via immersive 360° imagery—similar to a first-person video game—allowing for previously infeasible tasks, such as virtually previewing routes, visually inspecting destinations, getting a "feel" for a neighborhood, or even remotely visiting world-class tourist locations such as the Grand Canyon [53], Petra's Treasury, or the Eiffel Tower in France.

In 2007, Google launched the first broadly available streetscape mapping tool called *Street View* (GSV), which uses instrumented cars to take geo-located 360° imagery uploaded to and navigable via Google Maps [50]. Now, every major mapping service from Meta *Mapillary* [89] to Apple *Look Around* [9] and Bing *Streetside* [91] provide interactive streetscape imagery. GSV alone has grown to over 220 billion images spanning 10 million miles across 100 countries [43] while Mapillary has over 2 billion crowdsourced panoramas in 140 countries [87]. Though widespread and growing, these streetscape tools remain fundamentally inaccessible to blind users due to their reliance on panoramic imagery, lack of image descriptions, and inaccessible controls.

With advances in street scene analysis [19, 25] and multimodal AI models [56, 85, 114], we now have an opportunity to redefine the experience of streetscape maps for blind users. We introduce *StreetViewAI*, a new, accessible street view experience using context-aware, real-time AI and accessible navigation controls (see Figure 1 and video). With StreetViewAI, users can interactively pan and move between panoramic images, learn about nearby roads, intersections, and places, hear real-time AI descriptions, and dynamically converse with a live, multimodal AI agent about the scene and local geography (Figure 2). StreetViewAI was designed iteratively, drawing on the experiences of two blind team members, feedback from professional designers and engineers in mixed-reality and mapping tools, and literature in accessible first-person gaming [11, 101, 122], mixed-reality [17, 55], and mapping [90, 93].

To evaluate StreetViewAI and investigate the potential of accessible streetscape tools, we recruited eleven blind participants—all screen reader users—for a 90-min in-person lab study. Participants engaged in a semi-structured interview about their use of navigation tools, maps, and AI before completing a tutorial of StreetViewAI and two sets of tasks: point-of-interest (POI) investigations and open-world navigation. Sessions completed with a brief reflection and ideas for future prototypes. From StreetViewAI logs, observational notes, and post-task interview responses, we analyzed *what* questions blind participants ask about streetscape imagery, *how* well our multimodal AI models performed, *how* participants navigated and oriented themselves in the virtual-physical world, and *how* participants felt about our tool, key concerns, and opportunities for improvement.

We found that users were able to effectively traverse street view images and navigate the virtual-physical world, primarily by listening to auto-generated contextual status messages (*e.g., "Central Perk Cafe is now to your left 16 meters away"*) and utilizing the multimodal AI. Interestingly, while participants could trigger AI-generated scene descriptions of their current view at any time, most preferred to interact with the context-aware *AI Chat Agent* where they could ask context-based questions about the current scene, previous views, and the local geography. Of the 816 questions asked in AI Chat, 703 (86.3%) were correctly answered, 32 were incorrect (3.9%), and the remaining were either partially correct (26; 3.2%) or the AI refused to answer (54; 6.6%). Our findings also highlight challenges related to mental models of pedestrian navigation *vs.* inherent limitations of streetscape imagery, a bias towards trusting AI output (even when wrong), and interaction design difficulties in creating concise audio feedback that appropriately combines complex information from navigation data, panoramic imagery, and spatial relationships.

As the first accessible street view tool, our work fundamentally advances research in accessible maps and contributes design guidelines for making immersive 360° imagery accessible, new interaction techniques to accessibly converse with a context-aware AI agent about street scenes, and enumerates new opportunities for the next-generation street view tools that use AI.

2 Related Work

We situate our work in research on accessible maps, mixed reality (XR), AI-generated image descriptions, and street scene analysis.

2.1 Accessible Maps

Making maps accessible to people who are blind or low-vision (BLV) is a grand challenge in HCI and accessibility [16, 34]. To translate map data—traditionally based on visual representations—into more accessible non-visual formats, researchers have explored a variety of techniques including audio descriptions [24, 41, 99], sonification (non-speech audio) [28, 64, 107], and tactile representations ranging from raised-line graphics [134] and 3D-printed artifacts [26, 65–67] to refreshable tactile displays [71, 103] and dynamic haptics using vibromotors [68, 130] or force-feedback input controllers [116]. Despite significant advancements, prior work has not explored streetscape tool accessibility, our focus.

Beyond traditional maps, researchers have investigated real-world navigation support [36, 76, 105, 124], such as maintaining orientation during travel [36], navigating the "last few meters" to a destination [105], or avoiding obstacles [124]. Others have developed and studied custom guidance systems such as *SoundScape* [90], *FootNotes* [38], and *NavCog* [2, 106]—all which combine real-time location sensing with spatial audio and/or audio guidance. Much of this research has matured into popular commercial tools or features, such as *"Voice Guidance"* in Google Maps [41] or beacon-based audio navigation in *BlindSquare* [93], *GoodMaps* [40], and *VoiceVista* [73]. While StreetViewAI is not an *in situ* navigation tool, it is designed to support route planning and open exploration before travel—in ways not previously possible (*e.g.,* finding business entrances *a priori*)—and should synergistically support these *in situ* tools.

StreetViewAI integrates real-time AI to support image-based analysis of street scenes and geographic queries, which connects to emerging work combining geographic information systems (GIS) with LLMs. For example, *ChatGeoAI* [86] and *GIS Copilot* [4] enable novice analysts to perform complex geospatial operations using LLMs. Relatedly, though not backed by LLMs, prior work has explored making geographic-based query and visualization systems accessible [83, 109, 115, 133]. For example, *Atlas.txt* introduced a novel data-to-text natural language generation system to communicate geo-referenced information through screen readers; however, it did not support interactive explorations. Most relevant to our work, *VoxLens* [109] enables screen reader users to verbally query geospatial data—for example, to understand COVID-19 cases by US state—and Reinders *et al.'s* recent Wizard-of-Oz (WoZ) study examining refreshable displays with WoZ'd conversational agents [103]. In both works, the authors examine and taxonomize the types of questions and information BLV users are interested in with accessible geoanalytics. We pursue a similar analysis of user questions but applied to streetscape imagery. Moreover, while promising, the above tools are focused on traditional 2D map-based representations rather than the integration of geospatial data, immersive imagery, and virtual navigation as we are here.

2.2 Accessible XR and Virtual Environments

StreetViewAI does not strictly fit into traditional definitions of Augmented Reality (AR) or Virtual Reality (VR) [111] but its immersive, first-person perspective and interactive navigation share characteristics with both, and we draw on emerging mixed-reality (XR) accessibility guidelines [32, 95]. Towards AR accessibility, research focuses on enhancing access to the *physical world* through interactive computing, including via sensory substitution techniques [27, 128], sonification [1, 88, 125], and, most recently, real-time object recognition and description using smartphone cameras [7, 15, 17, 20, 39, 92] or AR glasses [58, 81, 82, 138]. For example, *WorldScribe* [17] uses emerging vision-language models (VLMs) to provide customized, succinct descriptions of a BLV's user's smartphone camera view. While related, StreetViewAI differs by operating on pre-captured 360° imagery rather than live sensor data. Moreover, StreetViewAI's interaction experience is disembodied, which introduces unique HCI challenges related to virtual locomotion, navigation, and views—while trying to maintain a spatial mental model of the world.

In this way, StreetViewAI is most related to research in accessible VR, which attempts to make virtual environments perceivable and navigable without vision, including via accessible input mechanisms like instrumented canes [110, 136] or echolocation [5] as well as output modalities such as spatialized audio and haptic feedback [108, 117, 120]. One early influential work includes *BlindAid* [108], a VR system using haptics and spatial audio to help BLV users learn layouts for Orientation and Mobility (O&M) training, sharing StreetViewAI's goal of virtual exploration for environmental understanding. More recently, researchers have applied real-time AI techniques in VR to improve accessibility [18, 23, 137], much as we do with StreetViewAI. For example, *EnVisionVR* [18] uses a VLM to support real-time scene descriptions and virtual object interaction; however, this is with purely *virtual* scenes rather than physical world data projected into an immersive environment.

Finally, we draw on the long history of accessible 3D first-person video games such as *Shades of Doom* [35], *AudioQuake* [11], and *Terraformers* [122], which demonstrate the feasibility of complex 3D navigation and interaction using keyboard interaction and spatialized sound. While StreetViewAI does not currently include spatialized audio, it borrows keyboard input mechanics for panning and movement common in these games (*e.g.,* arrows for movement).

2.3 Image Accessibility and AI Descriptions

Streetscape tools are fundamentally reliant on visual imagery, posing a significant accessibility barrier. Traditionally, image accessibility has depended on manually authored alternative text (alt-text) [126]. However, alt-text is frequently missing [37, 123], often inadequate for conveying the rich detail of complex scenes like streetscapes, and impractical to manually create for the billions of images hosted by services like GSV and Mapillary [43, 87]. This necessitates automated approaches. Recent advances in AI, particularly LLMs and VLMs, have enabled powerful new image

Figure 3: We iteratively designed StreetViewAI with two blind collaborators, improving the navigation experience, integrating additional geographic knowledge bases (*e.g.,* nearby places, intersections, *etc.*), adding in an AI chat interface, implementing additional hot key support, and refining the AI prompts. For legibility, the on-screen text size was artificially increased in these screenshots—text is natively voiced by StreetViewAI or read by the screen reader.

description techniques [29, 59, 79, 127, 135] and can support *Visual Question Answering* (VQA) systems where users ask specific questions about images [6]. These technologies are increasingly integrated into mainstream platforms—both Apple *VoiceOver* [8] and Android *TalkBack* [42] support AI-generated image descriptions—as well as tools like *BeMyAI* [15] and *Seeing AI* [92]. While important progress, these AI-based approaches are built to support general image-based descriptions and lack geographic context, viewport history, and support for virtual inquiry, as we do with StreetViewAI.

Recognizing the limitations of static alt-text descriptions, others have also explored interactive approaches such as *ImageExplorer* [79] and *AI Vision* [135], which provide hierarchical descriptions and detailed object exploration *within* an image. While enhancing interaction with a single image, these systems are not designed for navigating *between* immersive scenes or grounding interactions within a dynamically changing virtual viewpoint. StreetViewAI advances this area by uniquely integrating context-aware AI capabilities directly within a dynamic, navigable, 3D spatial interface.

2.4 Street Scene Analysis

Finally, StreetViewAI's ability to provide meaningful descriptions relies on advancements in AI-based street scene analysis and understanding [19, 25, 98, 139]. Most relevant to our work is research in automatically identifying pedestrian-related and accessibility features, including curb ramps [62, 63], sidewalk presence or condition [70, 121], specific accessibility obstacles [30], sidewalk surface materials [69], traffic signs [12], and assessing overall streetscape walkability [72, 80]. Perhaps most related is *BusStopCV* [78], a web app with an embedded *YOLOv8* model [102] that analyzes the user's streetscape view in real-time to help identify and collect accessibility information about bus stops (*e.g.,* presence of benches, shelters). Though not aimed at screen reader accessibility, BusStopCV demonstrates the potential of automatically extracting information from GSV in real-time as the user interacts with a street view scene.

Together, the above work demonstrates the power of AI-based analysis of street view imagery at scale; however, the focus is either on CV model improvements or efficient data acquisition for urban planning and mapping databases (*e.g.,* [70]). In contrast, our goal is to leverage these emerging techniques to provide real-time, interactive, and conversational access to the visual information within streetscapes. We do not aim to advance the state-of-the-art in CV models themselves but rather to apply these continuously improving models to create a novel accessible mapping experience.

3 Designing StreetViewAI

To design and build StreetViewAI, we followed an iterative, human-centered design process that included nine co-design sessions with our two blind collaborators (CB1, CB2)—who are also coauthors, feedback from professional designers and engineers with expertise in maps and AI-based mixed-reality, and drawing on our own decades of experience in creating accessible tools. Co-design sessions began with initial ideation and use of an early-stage prototype, which was progressively improved through feedback (Figure 3). We synthesize key findings and design considerations below.

Initial reactions. Though both of our blind collaborators were aware of streetscape tools, they had never used them due to inaccessibility. As CB2 described, *"I have heard of [Street View] but I couldn't use it since it is not accessible."* After using the StreetViewAI prototype for the first time, CB1 responded *"I think there is massive value here"* and CB2 said, *"This is so exciting, just having a conversation about the street [with AI] is really amazing."* Both co-designers emphasized the importance of user agency and freedom—the ability to explore, encounter features of interest, and drill down: *"I just want to learn more about what is around. I want to explore. And part of exploration is serendipity and the ability to see something and investigate more"* (CB1).

Learning mental models of street view. After some initial sessions, we realized that the concept of street view itself was unclear; thus, we worked to improve our explanation of *what* street view is (the concept of immersive and fully interactive 360° imagery), *how*, *where*, and *when* the images are taken (from the street by instrumented vehicles with 360° cameras), and *how* the immersive images are distributed and connected (available every 10-15 meters). Moreover, the inherent limitations of street view were unclear: how *close* could you get to buildings? Could you go *inside* buildings?

Navigation. For navigation, both co-designers emphasized wanting pertinent information on every "virtual" step (movement from one pano to another), including street addresses, nearby places, distance moved, and whether they had previously visited a location. Both worried, however, about verbosity, *"The information is* *really great but the verbosity will have to be trimmed down"*, and suggested allowing the user to control what they wanted to hear and adding non-verbal auditory signals (*i.e.,* earcons) for certain events. Moreover, both suggested additional movement controls beyond heading changes and virtual steps. For example, when virtually navigating from his apartment to a local shopping mall with StreetViewAI—along a route he knows well—CB1 suggested adding a "jump" feature to move by some larger increment or to a proximal landmark along the street (*e.g.,* an intersection).

Orientation. Orienting oneself geographically while maintaining notions of heading, relative position, and nearby places can be challenging and becomes even more complicated with virtual movement (where actions are disembodied). Our co-designers suggested adding quick access to key shortcuts for *"Where am I?"*, *"What am I facing?"*, *"What road am I on?"*, and *"What is the next intersection?"*.

Use of AI. The AI chat interface seemed to gain increasing importance as we iterated on the tool. As CB2 stated, *"The chat tool is so important; once you get to a destination, you can learn so many things like, how busy the street is? What are the nearby intersections? What does the door look like? Where is the door?"*. Beyond conversations about the scene, our co-designers suggested supporting commands to use the AI to direct movement or view changes: *"Can you ask the AI chat to move you or turn you around? Or to take you places: 'take me to the nearest ice cream shop?"* Finally, while we initially auto-generated an AI description upon every view change (either from panning or movement), to reduce information overload, we ultimately decided to require an explicit hotkey press for AI descriptions (Alt + D) and the AI chat (Alt + C).

3.1 Design Considerations

Drawing on our co-design sessions and key literature, we synthesized the following design considerations:

Concision. Between the street view image itself and the surrounding geography (*e.g.,* roads, intersections, nearby places), the streetscape navigation experience is information dense. A key design challenge then is providing high value, pertinent information with every viewpoint change: concise and clear.

Support conversational exploration. Allow the user to easily engage with the AI chat to quickly ask questions and follow-ups about the scene and location. Chat questions are inherently personalized to the user's interest in the moment: *"Does the sidewalk look shaded?"*, *"Is the entrance to the coffee shop wheelchair accessible?"*, *"Is there anything surprising along this route?"*.

Leverage streetscape images. At every interaction, harvest as much as possible from the street view image relevant to a blind or low-vision pedestrian, including safety features, sidewalk obstacles, and ambient qualities like shade. Go beyond what is currently available in existing map APIs (*e.g.,* in Google Maps or OpenStreetMaps).

Support "last 10 meters" navigation. Traditional mapping tools support origin-to-destination routing but only at the building-level granularity. We aim to support "last 10 meters" navigation [105], including the accessibility of sidewalks in front of buildings and the location of entrances.

Enable personalization. Allow users to tell the AI about themselves and their needs (*e.g.,* mobility aid usage, preferred verbosity level, types of points-of-interests).

Key use cases. Finally, we distilled the following key use cases: (1) point-of-interest (POI) investigation—*"What does a destination look like? Where is the entrance?"* (2) open-world navigation—*"I want the freedom to explore a region"*; (3) route planning—*"What might my walking route look like from A to B?"*; and virtual tourism—*"I've always wanted to visit Tokyo, Sydney, or even the Grand Canyon."*

4 The StreetViewAI System

Informed by our co-design sessions and prior literature, we built and evaluated the final StreetViewAI prototype, the first interactive streetscape mapping tool for BLV users, which uses a context-aware, multimodal AI to describe scenes (Figure 1) and accessible navigation controls. With StreetViewAI, users can virtually navigate the physical world via keyboard interactions, hear AI descriptions of the current scene, or chat with a context-aware, multimodal AI agent. See the video demo. Below, we describe key components, including panning and movement, the AI subsystem, and other accessible controls. We built the app to be self-voicing, though this can be toggled on/off to work with screen readers. For the latter, we use ARIA live regions (HTML elements with `aria-live`) to provide status messages and AI output to a screen reader.

4.1 Panning and Movement

We describe the user actions and underlying algorithms to support keyboard-based, non-visual interaction with StreetViewAI. As they interact, the user can trigger an AI description of their current view with Alt + D or chat with the AI agent via Alt + C (typing) or Alt + Spacebar (speaking). For example, after turning 90° or taking a step, the user can hit Alt + D to hear an AI-generated summary of the current view. Specific thresholds mentioned below were developed iteratively through internal testing and our co-design sessions (but are parameterized and easily changeable).

Panning. Unlike traditional 2D images, street view images are high-resolution panoramas—140 megapixels or 16,733 x 8,366 pixels [77]—spherically projected into an immersive, interactive scene

Key	Action
Movement Controls	
← →	Rotate left or right 45°
↑ ↓	Move forward/backward at current heading (if possible)
Alt + B	Go **b**ack to the last location (pano)
Alt + J	**J**ump to the next intersection or 70 meters (230 feet), whichever is first at your current heading
AI Interaction	
Alt + D	**D**escribe the current view with AI
Alt + C	**C**hat with the AI agent (typing)
Alt + Spacebar	Talk with the AI agent (speaking)
Alt + A	Repeat the previous output **a**gain
Esc	Stop the current speech output
Location Information	
Alt + W	**W**here am I? Get current address and heading
Alt + N	Get information about **n**earby places
Alt + I	Get information about the current and next **i**ntersection (if any)
Alt + M	Get possible **m**ovements at current location
Meta Information	
Alt + V	Get your **v**isit history for current pano
Alt + P	Get the date and photographer of current pano (**p**hotograph)

Table 1: List of hotkeys and their associated behaviors.

Figure 4: *AI Describer* uses one of two prompts: (a) a "default prompt" emphasizing navigation and safety for BLV pedestrians and (b) a "tour guide" prompt that additionally emphasizes tourism information such as historic and architectural context. (c) The MLLM model is fed the custom prompt along with an optional user profile and dynamic, context-aware input such as the user's location, nearby places, and their current view. The output is structured with (d) mobility features, obstacles, and a safety summary along with the (e) default or (f) tourist description, tailored from the prompt personality.

(similar to a VR world). The GSV API provides decimal-level control of the user's point-of-view (POV) with 360° horizontal and 180° vertical panning. To simplify interaction and limit disorientation, we discretize the horizontal (heading) panning space into octants (45° increments) and fix the vertical panning (pitch) to 0° (so, the user is looking "straight out" roughly at eye level). The user can pan left and right 45° using the ← → arrow keys, respectively.

As the user pans, StreetViewAI immediately voices the current heading as a cardinal or intercardinal direction (*e.g., "Now facing: North"* or *"Northeast"*), expresses whether the user can move forward along that heading and, if so, to what road address, and also explains whether the user is now facing a nearby place. To generate the places description, we maintain a list of nearby places at each panorama (pano) location using the *Google Places API* [47]. When the user's heading shifts, we describe places in front of the user—within a 45° angle of the current heading and a maximum distance threshold of 35 meters.

Taking a step. If a nearby pano is available at the current heading, the user can take a virtual "step" forward using the ↑ arrow. Similarly, if a pano is available opposite their current heading (180°), they can step backwards with ↓ (but will still maintain their current forward heading). These steps are approximately 5-15 meters depending on the GSV pano distribution in that geographic region.

To calculate navigable panos reachable by a "step", we derived our own custom algorithm suitable for non-visual interaction. Each `StreetViewPanorama` object in the GSV API maintains a list of connected panos; these `StreetViewLinks` include a heading (oriented from the current pano), a description, and a unique 'pano id'. Thus, one can navigate in GSV simply by traversing this node-link network (drawn as blue circles and edges in Figure 5); however, Google does not publish how they decide which panos to connect and, in practice, we found this network insufficient to enable non-visual navigation. For example, a user could be in the middle of a four-way intersection with a North-South street and an East-West street but at a pano with only two links, which artificially limits movement and causes confusion (see examples in Figure 5). The problem is: the road network does not match the underlying GSV network.

Thus, for each location, we construct a custom egocentric GSV graph, which better matches user navigability expectations and the underlying road network. For this, we perform a parallel search of

Figure 5: To support accessible streetscape navigation, we construct a custom egocentric GSV graph. From the user's current location (orange circle with heading), connected `StreetViewLinks` are drawn with edges and shown in blue; the current "move forward" pano is shown in green. (a) In this intersection, the built-in GSV graph is fine—the user can appropriately move north, south, east, and west; (b) In contrast, here there are only two built-in links restricting travel to east-west; our custom graph addresses this limitation and still allows for north, south, east, west travel, matching the intersection. (c) Similarly, this four-way intersection only has two built-in GSV links (two directions of travel) while our solution allows for all four, again matching the roads.

nearby panoramas along a 20x20 meter search grid with 5 meter step increments. We identify all panos within the grid but also all `StreetViewLink` panos (if not in the nearby list). For each found pano, we compute a distance and relative heading to the user's current location. With this data structure, we can immediately calculate the next and previous panos as the user changes their heading. If multiple panos exist along the same trajectory, we default to the closest one. If the user attempts to move forward or backward but no pano exists along that heading, we voice an *available movements* status message (also triggerable by Alt + M), which expresses possible movements at the user's current location. For example, *"You cannot move [forward/backward] along your current heading of East. You can move in three directions: Northwest, West, and Southeast."*

After taking a step, StreetViewAI voices the movement type (*e.g.,* step, jump), how far the user traveled, their current address, whether they moved to a new road, whether they have just arrived or left an intersection, a list of nearby places (within 50 meters) prioritized by direction, and whether they have previously visited the specific pano. For brevity, we only voice information that changes from the previous pano. Moreover, the list of nearby places is contextual and uses relative positions. For example, *"Starbucks Coffee is now on your left 12 meters away"* or *"The Talking Book and Braille Library is still in front of you but now 32 meters away."*

Jumping. In addition to "stepping", the user can also *jump* along their current heading by 70 meters (230 feet) or to the nearest intersection, whichever comes first. Because the Google Maps API does not provide road intersection data, we implemented our own detection algorithm. Our method casts a ray from the user's current location projected along their current heading and incrementally checks each point for an intersection with a specified step size (15 meters), search grid size (20x20 meters), and maximum distance (70 meters). To detect intersections at each step, we use the *Google Roads API* [48] to request roads and their geometry within the search grid. If two or more roads are found, we then run a parallel search for nearby panos within the same grid and determine if at least one pano has adjacent neighbors with assigned addresses of different roads. This works well in practice but does mean that bridges where two roads cross but do not intersect are treated the same as actual intersecting roads. After jumping, we generate a similar status message as a "step" but tell the user they "jumped."

Teleporting. Finally, the user can teleport by searching for and selecting a specific place or address in the "search box" field. This functionality is built on top of the `searchByText` endpoint in the Google Places API. As the user types, StreetViewAI returns a list of matching results in real-time; the list includes a display name, address, location type, and brief editorial summary. Once a new destination is selected, we virtually "teleport" the user to the closest nearby pano. Crucially, to help orient the user upon landing, we automatically adjust the POV to point to the destination.

Once the user teleports, we generate a more verbose status message to orient the user to their new location, including announcing the city, state, and/or country if they changed. For example:

> *You teleported 2,393 km from the Acropolis of Athens to 38 Bankside in London, England. There are four places within 50 meters, including: a performing arts theater, Shakespeare's Globe, is ahead of you 26 meters away; a historical landmark, Sam Wanamaker Plaque and an event venue, Ingresso, are to your left less than 47 meters away, and a garden, Tate Community Garden, are to your right less than 49 meters away. You are facing South and can move in four directions: South, West, Northwest, and East.*

Go back. If the user made a mistake or simply wants to return to their most recently visited pano, they can hit the Alt + B keys. This undo stack is currently of size one but is parameterizable for future iterations of the tool.

4.2 StreetViewAI's Multimodal AI

At any point, the user can trigger an AI-based analysis of their current view to learn more information. StreetViewAI has three AI subsystems, all built on top of Google's *Gemini Flash 2.0* model [46], which was selected due to its benchmark performance, ease of prototyping, and multimodal support [56]. The three AI subsystems include: *AI Describer*, *AI Chat*, and *AI Tour Guide*, each which use custom prompts informed by prior work [17, 20, 21], our co-design sessions, and iterated upon through internal testing. Unlike traditional AI-based image description tools [79, 92, 127], our AI models are multimodal and can consider relevant contextual information about the user, surrounding geography, and the image itself (Figures 2, 4). We describe each input component before the AI subsystems.

User profile. Users can optionally set a user profile, describing their vision level, mobility needs, and other relevant information, which is included as context in the prompt. For example, *"I lost my vision at age 19 due to Stargardt disease. For navigation, I use a service dog and guided audio directions in Google Maps. For travel, I prefer to walk, especially if the distances are under one mile."* If this profile is not set, the prompt asks the model to *"assume the user is blind and may use a white cane or a guide dog for mobility."*

Geographic context. We also assemble a dynamic list of geographic elements nearby the user, including the user's currently selected place (if any), their closest address, current heading, neighborhood, city, state/province, country, and a sorted list of nearby places with their business names, types, editorial summaries, `<lat,lng>` positions, distances from the user, heading offsets, relative positions (*e.g.,* left or behind the user).

User's view. Finally, we upload a 640x640 image of the user's current view, extracted from their current position and heading. Below, we describe each of the three AI subsystems. See supplementary materials for the full prompts.

Feature	AI Describer	AI Chat Agent	AI Tour Guide
Primary Goal	Concise scene awareness for orientation & navigation	Navigational cues, Interactive Q&A, System control	Engaging, informative virtual tour experience
AI Persona	Expert scene describer for BLV users	Helpful, responsive conversational agent	Knowledgeable, engaging tour guide for BLV tourists
Key Focus	Key objects, spatial relationships, navigational elements	Answering questions, executing commands	Historical facts, cultural significance, points of interest
Output Style	Short (2-3 sentences) or JSON	Conversational text or speech	Descriptive narrative (4-5 sentences)
Interactivity	On-demand description generation	Real-time chat with conversation memory	On-demand enriched description
Full Prompt	Supplementary Material S1.A	Supplementary Material S1.B	Supplementary Material S1.C

Table 2: Comparison of the three AI subsystems highlighting their distinct goals, personas, and functionalities.

4.2.1 *AI Describer.*

To invoke the *AI Describer*, the user can press Alt + D, which combines the three data components above within a custom prompt that begins:

> *You are an expert in describing visual scenes for people who are blind or have low vision. You will be assisting users who are navigating Google Street View using a screen reader. They rely on detailed audio descriptions to understand their surroundings. Your task is to describe this image concisely and accurately, capturing key details, as if you are describing it to a person who cannot see it.*

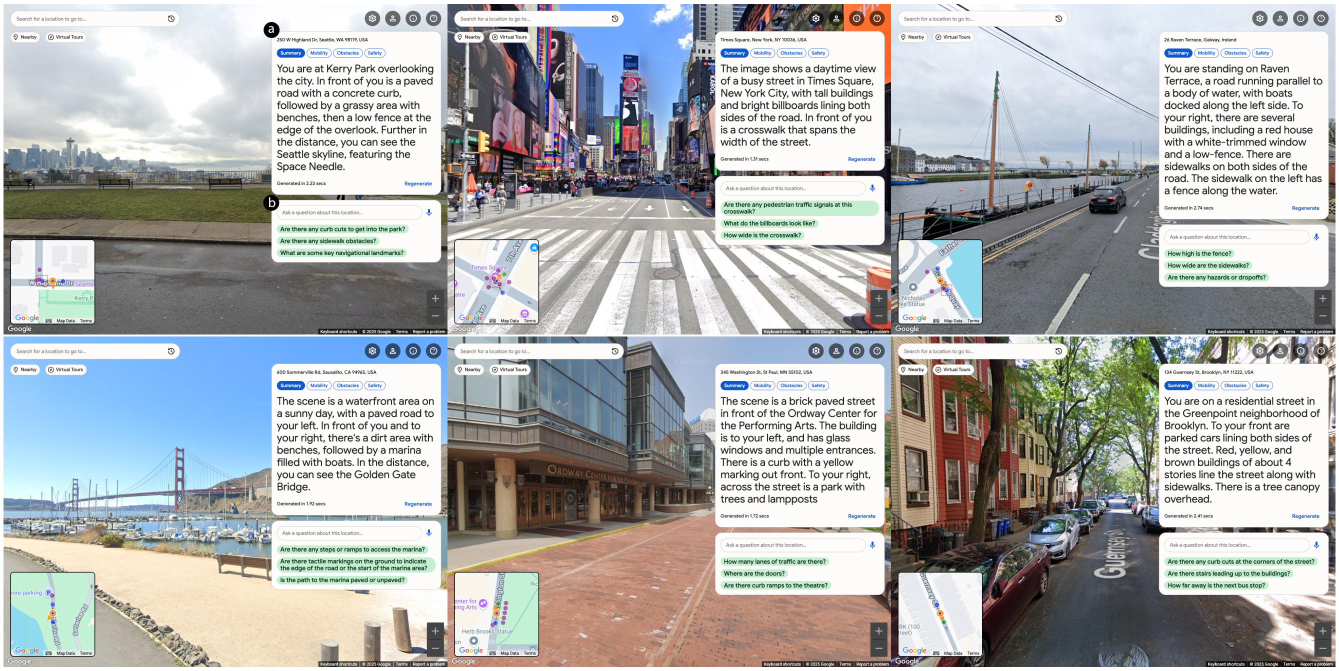


Figure 6: Example output from StreetViewAI's *AI Describer* across diverse urban and residential scenes. (a) The MLLM generated descriptions and (b) three context-relevant *AI Chat* suggestions for followup questions, enabling serendipitous discovery.

The prompt also tells the model to focus on eight specific areas, including key objects, spatial relationships, and navigational cues in the scene deemed relevant to a blind person and enumerates important guidelines such as using clear and concise language, a consistent frame of reference, to speak in the present tense, and to limit descriptions to two or three sentences. See Figure 4 for AI Describer's system diagram and Figure 6 for example scenes and output. In addition to Alt + D, the user can tab to and select the "Generate Description" button, which uses a similar prompt but requests a structured JSON response from Gemini with mobility-related features, potential navigation obstacles for BLV pedestrians, safety risks, and a list of three follow-up questions that the user could ask in AI Chat, which are appended to the UI (*e.g.,* Figure 6b).

4.2.2 AI Chat Agent. Complementing AI Describer, the *AI Chat Agent* allows for conversational interactions about the user's current and past views as well as nearby geography (Figures 2 and 7). The agent uses Google's *Multimodal Live API* [54], which supports real-time interaction, function calling, and, crucially, retains memory of all interactions within a single session. When the user initiates a chat either via typing (Alt + C), speaking (Alt + Spacebar), or both, we begin a WebSocket connection that persists throughout the chat. Every pan and movement interaction is then sent to the AI Chat Agent along with a corresponding image of the user's current view and geographic context (*e.g.,* nearby places, user's current heading, *etc.*). The context window is set to a maximum of 1,048,576 input tokens, which is roughly equivalent to over 4k input images or 8 average-length English novels.

Because of this multimodal context window, the user can ask contextualized questions such as *"Where is the bus stop?"* and, if the bus stop existed in previous views or in the geographic data, the agent can appropriately respond with a contextualized answer *"The bus stop is behind you, approximately 12 meters away."*—see Figure 9. While the Live API retains conversational, navigational, and view history, we did not specifically tune nor supply an explicit spatial orientation model. Though we observed frequently correct spatial interpretations in our user study, more work is necessary to investigate how to achieve robust spatial orientation in MLLMs, especially with dynamic movement and longer, more complex navigational histories (see 8.2 in the Discussion).

All AI Chat responses are spoken either natively by StreetViewAI using "text-to-speech" (if native voicing mode is turned on) or the user's screen reader via ARIA live messages. While both typing and speaking results in the same underlying function calls, in practice, we found that using speech input enabled a more seamless, multimodal input experience where the user could pan or move and immediately vocalize a question such as *"What am I looking at now?"* or *"Is there anything interesting here?"*. We use two audio earcons to signify when the mic is turned on/off.

Commands. In addition to asking questions, the user can also control StreetViewAI via commands spoken or typed to the agent, including to pan, move, and jump (*e.g.,* "turn left" or "move forward"; see Table 3). This is implemented via the Live API's `tools` method, which takes function declarations and descriptions as input. We parse the agent's response and call the appropriate functions. Even with these simple controls, we found that natural language commands were possible: instead of *"Turn me left by 45 degrees,"* the user could say *"Turn left a little bit."* Even further and more surprisingly, more complex command behavior emerged where the model incorporated context to enable semantic commands like *"Turn me towards the cafe."* See video demo and Figure 2c.

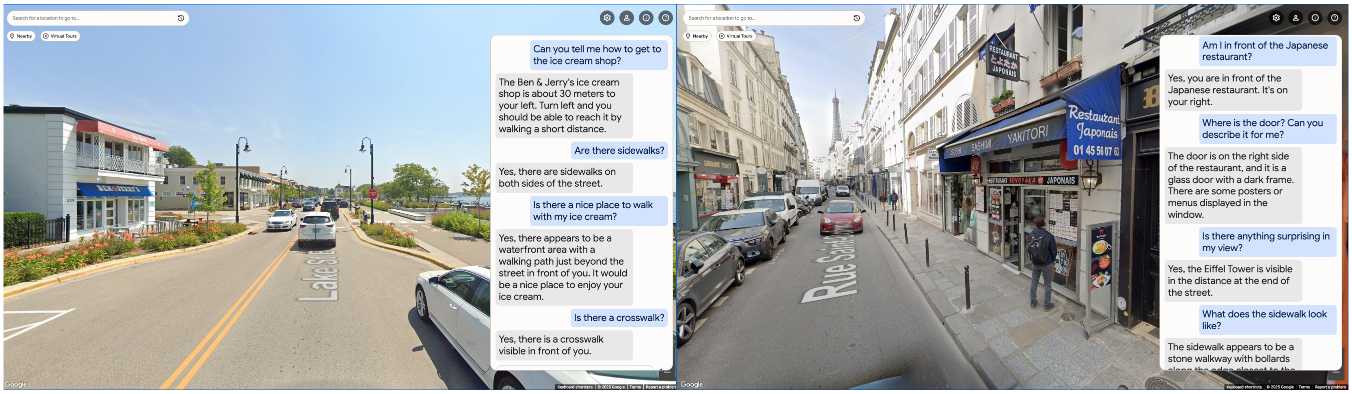


Figure 7: *AI Chat* examples drawn from the open-world navigation tasks in Part 3 of our user study. See also the video demo.

Function Name	Function Description
`moveBackward`	Move the user backward opposite the current heading to the previous Google Street View panorama
`moveForward`	Move the user forward along the current heading to the next Google Street View panorama
`moveToIntersection`	Move the user's location to the next intersection
`turnLeft45`	Turn the user's heading a bit to the left (-45 degrees)
`turnLeft90`	Turn the user's heading to the left (-90 degrees)
`turnRight45`	Turn the user's heading a bit to the right (45 degrees)
`turnRight90`	Turn the user's heading to the right (90 degrees)
`turnAround`	Turn the user's heading around (180 degrees)

Table 3: Function declarations and descriptions provided to the *AI Chat Agent* for controlling StreetViewAI via chat.

4.2.3 AI Tour Guide. Finally, the *AI Tour Guide* subsystem is essentially AI Describer but with a specialized prompt directing the Gemini model to act as an *"expert tour guide for blind or low-vision virtual tourists"* (Figure 4b). The prompt instructs the AI to combine clear visual descriptions with tourist-focused content, including historical facts, cultural significance, architectural styles, interesting anecdotes, nearby popular attractions, and descriptions of human activity in the scene to create a more engaging and informative experience, akin to a real guided tour (*e.g.,* Figure 8).

4.3 Other Controls

Beyond the commands already mentioned, the user can press Alt + W to hear their current address and heading, Alt + I for nearby intersections, Alt + N for nearby places, and Alt + P to hear the date and photographer of the current pano (*e.g., "This Street View image was taken on February 2025 by Google"*). See Table 1.

5 Implementation

StreetViewAI's frontend is implemented in TypeScript, Sass (SCSS), and HTML with the open-source libraries *Lit* for declarative UI components [84], *RxJS* for reactive state management [104], and *MobX* for efficient state observation and synchronization [94]. The backend is implemented in a custom lightweight *Python* server (similar to *Flask*), which serves static app resources. No data is currently stored on our web server—key functionality, including AI implementations, are done via client-side implementations and external API calls, all which use the Google Cloud infrastructure.

For nearby places, road information, and the interactive streetscape images themselves, we use the *Google Maps Places* [47], *Roads* [48], and *StreetView* [49] APIs, respectively. StreetViewAI is designed to be natively voiced—toggable to work with or without screen readers with a configurable speech output rate—which is implemented via Google Cloud's "*text-to-speech*" API [45]. The voice recognition and transcription is implemented with Google's "*speech-to-text*" API [44]. The AI subsystems were implemented with Google's *Vertex AI* platform and *Gemini 2.0 Flash Model* [51]. To more easily discern status update messages *vs.* AI Chat responses, status messages are read with a male voice ("en-US-Wavenet-J") and the chat with a female voice ("en-US-Wavenet-G")—see [52]. For efficiency, all API calls are cached in the browser using a custom, location-based *IndexedDB* instance.

6 User Study

To evaluate StreetViewAI and explore the potential of an accessible, AI-driven street view experience, we conducted an in-person lab study with eleven blind participants. The study tasks are informed by our co-design experiences as well as relevant prior work in BLV navigation, including our own (*e.g.,* SoundScape[90]). We focus on POI investigations and virtual route exploration.

6.1 Participants

We recruited eleven Participants (6 men, 5 women) using mailing lists to local blind and low-vision organizations, newsletter posts, and a contact list from previous studies. Participants ranged in age from 20–66+. All used white canes for mobility; two also used guide dogs. For technology, all participants used screen readers, predominantly *JAWS* [33] on *Windows* and *VoiceOver* on *iOS*. Most reported experience with at least one mapping tool, most commonly *Google* or *Apple Maps* but also blind-specific navigation tools such as *BlindSquare* [93], *VoiceVista* [73], and *GoodMaps* [40]. While technology experience and comfort with laptops (and typing) varied, seven participants had medium-to-high familiarity with AI tools, most commonly *Be My AI* [15], *ChatGPT*, and *Seeing AI* [92]. Three wore Meta AI glasses but indicated using the camera+AI features rarely and, instead, relying on the convenient headphone output for

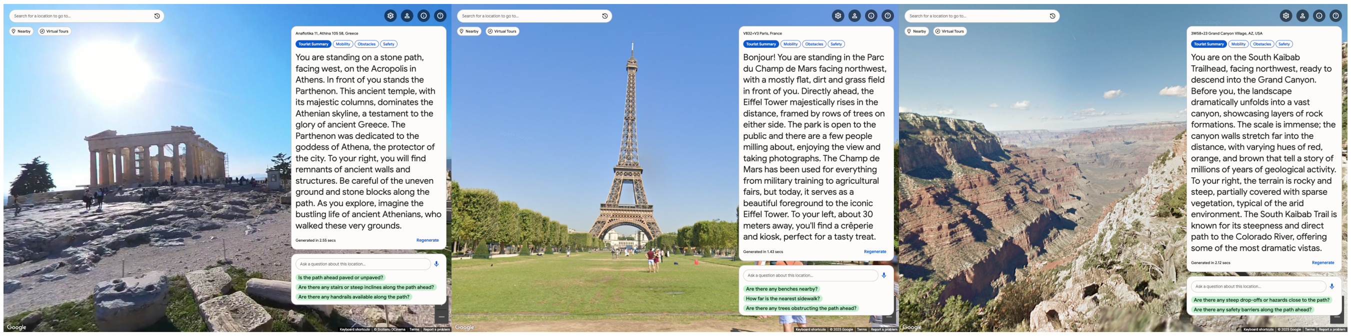


Figure 8: Virtually visiting the Parthenon, the Eiffel Tower, and the Grand Canyon using the *AI Tour Guide*. With StreetViewAI, users can virtually travel to over 220 billion street view images spanning 100 countries [43].

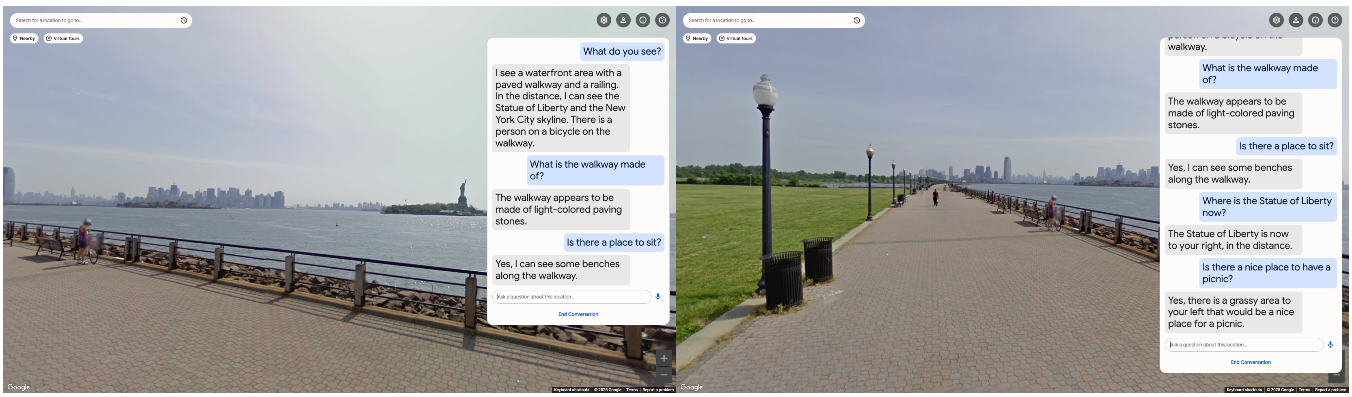


Figure 9: Demonstrating spatial orientation support in *AI Chat*. The user turned to the left in the second image.

their screen readers. Though a few participants had heard of GSV, none reported experience with any streetscape tools, highlighting an opportunity for this population.

6.2 Procedure

The single-session, in-person lab study consisted of four parts and lasted 1.5-2 hrs: (1) a formative interview inquiring about the use of screen readers, navigation tools, maps, and AI; (2) a brief introduction and tutorial task with StreetViewAI followed by points-of-interest (POI) investigations; (3) two open-world navigation tasks in pre-selected areas followed by a third open-world navigation at a location of the participant's choice; (4) and a debrief interview. For our selected geographies, we specifically chose locations that would be unfamiliar to the participant so they would rely on StreetViewAI rather than their own memories. See the supplementary materials for the full protocol.

Because of the complexity and open-endedness of the study tasks, the varying technical proficiency of the participants, and because we wanted to observe users naturally engage with StreetViewAI and ask questions of the AI, not all participants made it to all tasks. The protocol was intentionally designed to accommodate different task completion speeds, and we skipped the last navigation task (as necessary). Prior to the study, participants filled out a demographic form asking about vision, screen reader usage, and assistive technology.

Part 1: Formative Interview. In Part 1, we asked participants about their use of screen readers and mapping tools, how they navigate physical environments, plan routes, and use accessible navigation tools for support (*e.g.,* BlindSquare) as well as their use of emerging AI tools (*e.g.,* Be My AI, ChatGPT). Finally, we discussed their familiarity and use, if any, of streetscape imagery tools.

Part 2: POI Investigations. Part 2 focused on using StreetViewAI to evaluate destinations (*i.e.,* points-of-interest or POIs). Our research questions included: *What* are BLV people interested in knowing about destinations? *What* questions do they ask? *How* well does our tool perform? We began with a brief introduction of what street view imagery is, how the data is collected, and a tutorial task where participants virtually visited a cafe in New York and learned basic controls, hot keys, and were asked to answer questions such as *"Are there chairs to sit on outside the cafe?"* and *"Are there obstacles on the sidewalk?"*. Participants were then invited to pan and move via the arrow keys and answer questions about the current intersection and other nearby places. We encouraged participants to ask their own questions and to be adventurous.

Following the tutorial, participants investigated three POIs with scenarios drawn from literature (*e.g.,* [60, 61, 74, 78, 113]) as well as our co-design sessions and, crucially, could only be addressed via visual inquiries with StreetViewAI: (1) investigating features at a bus stop; (2) assessing a playground for fun and safety, and (3) examining an unfamiliar Mexican restaurant and surroundings.

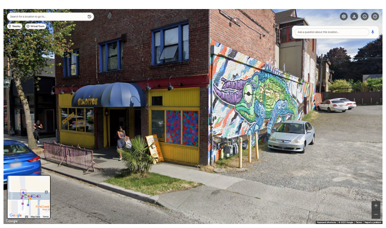
StreetView	Task	Questions
	**Tutorial task: Coffee Shop** We will virtually teleport to a coffee shop in New York called Irving Farm New York. Your goal is to evaluate the coffee shop and the surroundings to help you navigate there in real life.	• Are there chairs to sit on outside the café? • Are there obstacles on the sidewalk? • What road are you on? • Are we at an intersection? If so, what? • What else is nearby me? • When was the photograph taken? • <Ask your own question>
	**Bus Stop** You are visiting a new neighborhood and need to learn about a bus stop. Use StreetViewSR to interact with the scene and learn about the bus stop features that would help you travel there in real life (e.g., navigational features, landmarks, etc.).	• What does the bus stop look like? • Are there benches and a bus shelter? • What are some surrounding navigational landmarks? • Are there garbage cans? • What are some nearby buildings?
	**Playground** You are visiting your sister's family and want to bring your two young nieces to a new playground. Your goal is to learn about the playground before going. What kind of equipment does it have? Does it seem safe?	• What does the playground look like? • Are there slides and/or swings? • What kind of ground cover is there? • Are there benches to sit on? • Are there sidewalks? • Does the street look busy? • What kind of neighborhood does it look like? • What kind of houses are nearby?
	**Mexican Restaurant** You are meeting a friend at a Mexican restaurant unfamiliar to both of you. Your friend heard there was something interesting about the building and also asked about whether to drive or bike (*e.g.,* is there a parking lot nearby or a bike rack for their bike?).	• What does the building look like? • What is unique about the building? • Is there a parking lot nearby? • Is there a bike rack? • Is there a sidewalk in front of the restaurant?

Table 4: The Part 2 "POI Investigation" tasks and questions for participants. For each scenario, we invited participants to explore the location on their own and ask their own questions before providing our specific list of questions. If the participant did not immediately know the answer to a posed question, they were encouraged to use StreetViewAI to find out.

Following open exploration time, we asked participants about the scene. See Table 4.

Part 3: Open-world Navigation. In Part 3, participants completed open-world navigation tasks to virtually explore potential walking routes in two geographies that we selected—an ice cream shop in Minneapolis and a Japanese restaurant in Paris—and then in an area familiar to the participant (*e.g.,* their own neighborhood). This sort of pre-journey planning and virtual route rehearsal is a cornerstone of *Orientation and Mobility* (O&M) training for BLV individuals [13]. Here, our research questions included: *How* do BLV users combine multiple information sources to navigate? *What* visual features in a scene are they most interested in? *What* parts of StreetViewAI are most useful to supporting open-world navigation?

For the first navigation task, participants were told to *"virtually investigate a walk from your Airbnb rental to a local ice cream shop and also determine what else is nearby (e.g., is there a nice place to eat your ice cream)"* and for the second: *"You are visiting a friend in Paris and meeting at a Japanese restaurant. You are investigating the walk from the nearest bus stop to the restaurant."* See Figure 11. For both tasks, we explained *which* direction to start walking but neither how far nor the end addresses. Moreover, we specifically started the participants with a view facing away from the destination. Notably, these tasks were designed as open-world navigation rather than routing tasks because the system did not know the user's destination (we return to routing in the Discussion).

6.3 Data and Analysis

With participant consent, all sessions were audio and video recorded using both room- and screen-based recordings. To monitor tool usage, we built a lightweight, client-side logger that recorded user commands and AI interactions (including prompts, responses, and system actions) as timestamped JSON files. The lead researcher took notes throughout each session, observing navigation behaviors, user reactions, and marking when the AI provided an inaccurate response or refused to answer. These notes were corroborated *post hoc* via log analysis and a review of the video recordings. After

PID	Gen.	Age	Vision	Mobility Aid	Screenreader	Mapping Tools	AI Fam.	AI Tools	Streetscape Tools
1	M	18-25	Blind	White cane	JAWS/NVDA on Windows; VoiceOver on iOS	BlindSqaure, GoodMaps, Google Maps, Lazarillo, Voice Vista	High	BeMyAI, ChatGPT, Copilot, Gemini, Image Describers	Never used
2	W	56-65	Low vision	White cane	JAWS/NVDA; VoiceOver on iOS; Braille display	BlindSquare, Google and Apple Maps	High	ChatGPT, Meta Glasses, VoiceOver image description	Never used
3	M	18-25	Blind, some light detection	White cane	JAWS/NVDA; VoiceOver on iOS; TalkBack on Android	BlindSquare, Google Maps	High	Gemini, TalkBack image description	Never used
4	W	46-55	Low vision	White cane	JAWS; VoiceOver on iOS	Google/Apple Maps	Low	None	Never used
5	M	26-35	Blind	White cane	JAWS/NVDA, VoiceOver on iOS, Braille display	Google/Apple Maps, VoiceVista	High	Be My AI, Copilot, Seeing AI, VoiceOver image description	Never used
6	W	46-55	Blind	White cane	JAWS; VoiceOver on iOS	Google Maps (but rarely)	Low	None	Never used
7	W	66+	Blind, some light detection	White cane	JAWS/Fusion; VoiceOver on iOS	Google/Apple Maps	Low	Tried Be My AI in class	Never used
8	W	26-35	Blind, some light detection	White cane	JAWS; VoiceOver on iOS	BlindSquare, Google Maps with Voice Guidance	High	Be My AI, Seeing AI	Never used
9	M	26-35	Blind	White cane	JAWS; VoiceOver on iOS	Google Maps with Voice Guidance	High	AIRA (but not with AI), Be My AI, ChatGPT	Never used
10	M	66+	Blind, some light detection	White cane, guide dog	JAWS/Fusion; VoiceOver on iOS	BlindSquare, GoodMaps, VoiceVista	Medium	ChatGPT, Meta Glasses, Perplexity, Seeing AI	Never used
11	M	66+	Blind	White cane, guide dog	JAWS; VoiceOver on Mac and iOS	BlindSquare, GoodMaps, Google Maps, VoiceVista	Medium	AIRA (but not with AI), Be My AI, Copilot, Meta Glasses	Never used

Table 5: Participant demographics, mobility aids, and technology experience. *AI Fam.* = AI Familiarity.

West North East South

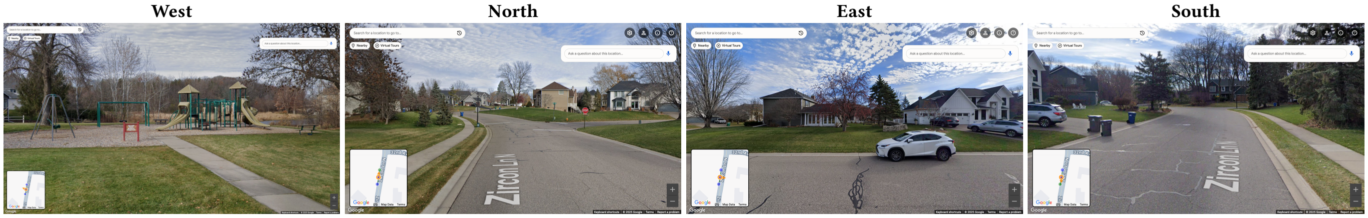


Figure 10: Directional views of the Part 2 "Playground Scenario". Unlike the other POI study tasks, the playground task required participants to pan beyond the initial view to understand the surrounding context (*e.g.,* residential neighborhood, quiet street).

each task, participants were asked to reflect on their experience and respond to 7-point Likert-scale questions on ease-of-use, information value, and perceived accuracy (7 is best). Participant responses were analyzed via thematic analysis [22].

Because study participants were using an unfamiliar keyboard and OS, we added tactile markers to the hot keys to improve usability. Still, five of the eleven participants inadvertently hit a key that closed StreetViewAI during the study: P8 once and P2, P5, P7, P10 twice each. Because our custom logger batches log data to persistent storage every ten entries, we consequently experienced mild data loss (~90 log entries total; < 0.6% of all logged events).

7 Study Findings

Overall, reacted positively to the prototype: *"Most navigation systems can get you to the last 5-10 feet but this helps you get to a door and even describes that door"* (P10), *"If I'm going to a place, I can familiarize myself first from my home"* (P1), and *"This is incredible!"* (P5). All eleven participants completed the POI investigations and 10/11 completed at least one open-world navigation task (P9 left before this study task began because his *Access Transit* arrived early). In total, participants moved to 356 panos, made 568 heading changes, and 1,053 AI requests (136 AI Describer and 917 AI Chats).

During the post-study debrief, participants rated the overall usefulness of StreetViewAI as 6.4/7 (*Median*=7; *SD*=0.9) and all wanted to use the system as a product, if available (*Avg*=6.6; *Med*=7; *SD*=0.8): *"Is this available? I want to use it now!"* (P10) and *"I'm very excited for this to come out, it's going to make a lot of blind people very happy"* (P5). While participants found value in StreetViewAI, relied heavily on its AI features, and adeptly combined virtual world navigation with AI interactions, they occasionally struggled with orientation, distinguishing the veracity of AI responses, and determining the limits of AI knowledge. Below, we first describe AI usage and performance followed by how participants accomplished the study tasks and high-level themes.

7.1 AI Usage and Technical Performance

Across the study tasks, participants used AI 1,053 times with a strong preference towards *AI Chat* (917 chats) *vs.* 136 *AI Describe* invocations[1]. On average, participants used AI Describe 12.4 times (*Median*=10; *SD*=7.5) *vs.* AI Chat 93.0 times (*Med=97.0*; *SD*=37.0) per session. As P10 said, *"The more you ask, the more you learn."*

[1]Participants had a total of 996 interactions with AI chat; however, 52 were inadvertent "live mic" instances (*e.g.,* where the system overheard a participant's "think-aloud" comment or a remark from the study facilitator) and 27 were misheard requests.

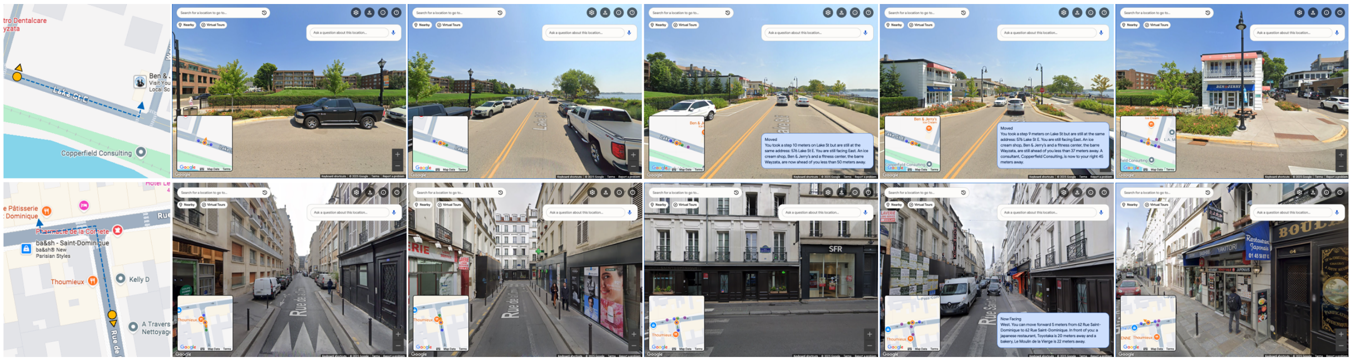


Figure 11: In Part 3, participants used StreetViewAI to complete two open-world navigation tasks: (a) to an ice cream shop, which required ~4 heading changes and ~12 steps (or 1 jump and 3 steps) and virtually traveling 107 meters (350 ft) (b) and to a Japanese restaurant in Paris, which required ~7 heading changes and ~5 steps (or 1 jump and 2 steps) across 73 meters (240 ft). On the left, the map shows the starting location and heading as an orange circle with the path to the destination in blue.

The AI response times were fast, on average less than a second (*Avg*=968.6 ms; *Med*=723; *SD*=678.8) with the AI Chat system—which uses Gemini's Live Multimodal API—responding faster (*Avg*=842 ms; *SD=614.3*) than AI Describe (*Avg*=1,916 ms; *SD=265.6*). Interestingly, a large majority of AI Chat questions were voiced (over 94.4%) perhaps due to a combination of participant unfamiliarity with the Mac keyboard, the flexibility and ease of voice, and the affordances of multimodal input (*i.e.,* keyboard + AI Chat simultaneously). Below, we analyze question types, AI response accuracy, and non-responses (*e.g., "I'm sorry, I cannot answer that."*). Quotes are from actual AI chat sessions in the study.

Question types. We analyzed all 917 AI Chat interactions and annotated each with up to three tags drawn from an emergent list of 23 question-type categories. Most commonly, with 27.0% of interactions (248/917), participants asked about spatial orientation of an object or themselves (*e.g., "How far is the bus stop from where I'm standing?"*, *"How close are the garbage cans to the benches?"*) followed closely by asking about object existence (243; 26.5%), including sidewalks, obstacles, and doors—*e.g., "Is there a merry-go-round here?"*, *"Are there soccer fields?"*, and *"Is there a bench at the park?"*. Third were questions asking for a description (169; 18.4%) such as "*What's in front of me?"* or *"What view do you see?"* followed by where an object or place is (137; 14.9%)—*e.g., "Where is the nearest intersection?"*, *"Where is the entrance to the restaurant?"*, or *"Can you help me find the door?"*. Demonstrating the conversational nature of AI Chat, over 14% (132) of chat interactions were followups such as *"Tell me more about the mural"* or *"Is the sidewalk connected to the path?"* or non-questions expressing gratitude or niceties (7.9%). Asking for specific directions (67; 7.3%) and about nearby places (85; 9.3%) was also common: *"How to get to the door without bumping into any chairs or tables?"* and *"Is there a train station close by?"*

Interestingly, over 6% of chats (59) were clarifications or confirmations (*e.g., "When you say the crosswalk is directly in front of me, what am I currently facing?"*, *"What do you mean by bollards?"*, or *"Just to confirm, there is no bike rack along the sidewalk?"*) as participants sought to strengthen their understanding or double check previous information. Many of the existence and location inquiries were related to safety or accessibility (51; 5.6%) (*e.g., "Is there an assistive pedestrian signal at this light?"*, *"Is there a marked crosswalk at this street?"*, *"Are there stairs along this route?"*), 30 (3.3%) were related to entrances (*e.g., "What does the door look like?"*, *"How do I get to the entrance?"*), and 15 (1.6%) were requests to read signage.

Question Type	Cnt	%	Example Questions
Spatial orientation	248	**27.0%**	*"How far is the bus stop from where I'm standing?", "How close are the garbage cans to the benches?", "What's to the left of this building?"*
Object existence	243	**26.5%**	*"Is there a merry-go-round here?", "Are there soccer fields?", "Is there a bench at the park?"*
Description requests	169	**18.4%**	*"What's in front of me?", "What view do you see?", "What do you see now?"*
Object/place location	137	**14.9%**	*"Where is the nearest intersection?", "Where is the restaurant entrance?", "Can you help me find the door?"*
Followup questions	132	**14.4%**	*"Tell me more about the mural", "Is the sidewalk connected to the path?"*
Nearby places	85	**9.3%**	*"Is there a train station close by?"*
Niceties/gratitude	73	**7.9%**	Expressions of gratitude or conversational niceties *"Thank you." "That's interesting!"*
Directions	67	**7.3%**	*"How do I get to the door without bumping into any chairs or tables?"*
Clarifications/confirmations	59	**6.4%**	*"When you say the crosswalk is directly in front of me, what am I currently facing?", "What do you mean by bollards?"*
Safety/accessibility	51	**5.6%**	*"Is there an assistive pedestrian signal at this light?", "Is there a marked crosswalk at this street?", "Are there stairs along this route?"*
Commands	40	**4.4%**	Moving, turning, jumping commands (*e.g., "Zoom into the bus stop sign to read the schedule"*)
Entrance-related	30	**3.3%**	*"What does the door look like?", "How do I get to the entrance?", "Are there obstacles to the door?"*
Additional info requests	29	**3.2%**	*"Can you tell me what's on the menu?", "What bus routes serve this stop?"*
State/condition inquiries	19	**2.1%**	*"Does this [street/sidewalk/restaurant] look busy?"*
Signage reading	15	**1.6%**	Requests to read signs or text

Table 6: Findings from our analysis of AI Chat across all study tasks (N=917). Each user question was annotated with up to three thematic tags drawn from an emergent list of 23 question-type categories.

Other unexpected but less common questions emerged such as asking about the state of something (19; 2.1%), such as *"Does this [street/sidewalk/restaurant] look busy?"*), asking questions that would require additional information sources beyond what we currently supply the model (29; 3.2%), such as *"Can you tell me what's on the menu?"* or *"What bus routes serve this stop?"*. Over 4% (40) of chat interactions were commands, including moving, turning, and jumping but not all of which the AI could perform. For example, P10 said: *"Zoom in to read the bus schedule"* in reaction to the AI explaining it could not read the sign because it was too far away; however, "zoom" is not currently implemented in StreetViewAI.

Accuracy. Of the 816 questions asked, 703 (86.3%) were correctly answered, 32 were incorrect (3.9%), 26 were partially correct (3.2%) and, for the the remaining 54 (6.6%), the AI told the user it could not provide an answer. Of the 32 incorrect responses, 20 (62.5%) were false negatives—*e.g.,* stating that a bike rack or park bench did not exist when it did, four (12.9%) were misidentifications (*e.g.,* a yellow speed bump interpreted as a crosswalk, a distant sign as a gate), and another four (12.9%) were errors that occurred when the user asked about the existence of something before the destination was fully in view (*e.g.,* asking if outdoor seating existed before being close enough to the restaurant to clearly see it). Though uncommon, two other interesting errors emerged: correctly identifying an object or place but misrepresenting its location (*e.g.,* the AI told one participant that the Japanese restaurant was on the corner but it is located on the middle of a block) and being told that you could move closer to a destination but you could not. Here, the AI model confused being physically present in a scene—where you could rightly walk forward to the destination but in StreetViewAI, due to the availability of GSV imagery, the participant was as close as possible. This understandably confused the participant.

Refusals. For the 54 queries that the AI refused to answer, it was most often because a request required geographic knowledge the AI did not have access to (25 of 54; 46.3%) such as asking about the nearest transit stop or getting directions to a place not in the immediate search radius (50 meters). Note that we did not connect the AI model to the full Google Places API (or other geographic knowledge sources), so its geographic data was limited. Similarly and second most commonly, participants asked questions that required additional knowledge bases such as available bus routes, transit schedules, or restaurant menus at particular locations (15/54; 27.8%). The remaining non-responses were due to unreadable signage, objects the AI couldn't see, or questions beyond the AI's capabilities. For example, one participant asked about relative distances between playground equipment, which resulted in the response *"I'm sorry, I cannot tell you exact distances between playground features"*, while another asked the AI agent to actually plan his visit to the playground with his nieces (*"I cannot plan your specific visit..."*).

We now turn towards describing the POI and open-world navigation tasks before enumerating some high-level themes.

7.2 POI Investigation

For each POI investigation, participants were told the task-specific scenario, teleported directly to the destination, and asked to explore on their own given the task (*e.g., "investigate the playground for safety and fun"*). At the end of each task, participants were asked questions about the location and told that if they did not immediately know the answer, they could use StreetViewAI to find out. To avoid influencing reactions to our prototype, the study facilitator did not reveal answer correctness until the debrief.

To accomplish the tasks, participants employed a variety of techniques but primarily relied on AI for information. Upon arrival, they listened carefully to the teleport message, which provided location data, nearby places, the user's heading, and possible movements. Most then began with an AI Describe description (Alt + D) before utilizing AI Chat (Alt + C) to drill down. Others simply started with chat. Overall, participants used AI Describe an average of 7.3 times per participant (*Med*=8.0; *SD*=3.4) and AI Chat 57.3 times (*Med*=55.0; *SD*=23.0) across the four locations (including the tutorial). While we instructed participants that it was not necessary to move or pan to complete the POI tasks (except for the playground task), most at least changed their heading at each location (*Avg=7.8*; *Med*=5), demonstrating an interest in looking around to gain more scene understanding. As P8 said:

> *This is literally what I do when I physically relocate to a new place or work at a new building, I try to build a whole mental model of everything that is around me so that I can orient myself. So, I think this would be very useful for that use case and, also, obviously for planning to go to unfamiliar places.*

Post-task questions. Impressively, all participants successfully answered the post-task questions with a high correctness rate (Table 7). Of the 20 scene-related questions, participants correctly answered 18.6 correctly (93%). Across all 220 questions asked (20 x 11 participants), there were 9 total incorrect answers, 4 partially correct, and 3 non-responses. All mistakes were due to inaccuracies in the AI and not due, for example, to participant misunderstanding. The most common AI error was stating that there was *not* a bike rack at the Mexican restaurant (there was, though it was not particularly salient), that there *was* outdoor seating at the restaurant (there wasn't), or that there were *not* benches at the park (there was but they were in the distance). Partial errors involved reporting only grass at the playground (there was also wood chips). Interestingly, as demonstrated by Table 7, the AI was reliable but randomly inconsistent across participants—responding accurately most of the time but in a small number of cases, failing (where it had previous succeeded) or omitting details. We return to accuracy perceptions in the themes below and in the Discussion.

7.3 Open-world Navigation

For the open-world navigation tasks, participants were first asked to find an ice cream shop in Minneapolis and then a Japanese restaurant in Paris (Figure 11). To accomplish the ice cream task, the participant needed to shift their view to the right (east) 90° and then walk (or jump) 350 feet to the first intersection. At the intersection, Ben and Jerry's was on the participant's left (north). For the Japanese restaurant task, the user needed to turn around (180°) to face north then walk or jump 130 feet (40m) to the nearest intersection and turn left (west), then walk 75 feet (23m) with the destination on the right. Once the user thought they had arrived at their destination, they were asked to answer questions relevant to the task (*e.g., "Is there a nice place to walk with their ice cream?"* or *"What food was advertised at the Japanese restaurant?"*).

Completing the tasks. All ten participants who began the ice cream task finished it but with varying strategies and efficiency. On average, it took participants 13.8 minutes with 10.2 movements, 28.3 heading changes, and 17.2 AI invocations (3.5 AI Describe; 13.7 AI Chat). The most successful participants listened intently to each status message (which occur at panning/moving), turned and virtually walked towards the ice cream shop until they heard "Ben and Jerry's" as a nearby place, and then continued stepping

Questions	Correct Response	Correct
Bus Stop		
What does the bus stop look like?	*"A covered bus stop with benches on a university campus"*	100%
Are there benches?	*"Yes"*	100%
Is there a bus shelter?	*"Yes"*	100%
What are some surrounding navigational landmarks?	*"Library, student union, bus stop shelter, and garbage cans to the left"*	100%
Are there garbage cans?	*"Yes"*	100%
What are some nearby buildings?	*"Student union, library, etc."*	100%
Playground		
How does the playground look?	*"A playground with swings, slides, and a climbing structure"*	100%
Are there slides?	*"Yes"*	100%
Are there swings?	*"Yes"*	100%
What kind of ground cover?	*"Grass with wood chips under the play structure"*	77.3%
Are there benches to sit on?	*"Yes"*	81.8%
Are there sidewalks?	*"Yes"*	100%
Does the street look busy?	*"No"*	100%
What kind of neighborhood?	*"Residential"*	100%
What kind of houses are nearby?	*"Two-story, single-family homes"*	100%
Mexican Restaurant		
What does the building look like?	*"2-story brick building w/a blue awning, yellow side, & a mural"*	77.3%
What's unique re: the building?	*"A mural. Large, green creature (dragon) with purple hair."*	100%
Is there a parking lot nearby?	*"Yes"*	90.9%
Is there a bike rack?	*"Yes"*	63.6%
Is there a sidewalk out front?	*"Yes"*	100%

Table 7: Post-task scene related questions and the correctness of participant answers. To address these questions, participants primarily relied on AI Chat but also AI Describe and the hotkey Alt + N for nearby places.

forward while asking the AI about the shop's specific location and whether it was visible. Those who struggled either walked or jumped past the ice cream shop and got confused or overly relied on nearby place descriptions (Alt + N) at a cost of involving AI Chat. Due to limited time, only 7 of the 11 participants conducted the Japanese task; all finished. Here, it took participants an average of 9.4 minutes with fewer movements and AI uses than the ice cream shop, perhaps because of the shorter distance or increased experience of the participants.

Own locations. Six participants had time to select and explore their own location; most chose their own homes or a favorite restaurant. One person selected a place where they studied abroad 25 years ago in England, another selected a home he grew up in, and a third explored the walk from their apartment to the *Department of Services of the Blind.* All participants enjoyed listening to the AI-generated descriptions, comparing them to their own understandings: P6 said, *"See, that's my house. And I could turn right here to get to 168th St."* (which she then did) and P1 noticed a difference, *"Oh, this image must have been taken when we still had a garage."* P11 visited *Chick-fil-A.* After teleporting and hearing the description, he instantly recognized that *"It took us to the wrong one"* as there were two Chick-fil-A's on the same street but a few miles apart. Similarly, P5 was able to discover that StreetViewAI brought him to the back of his father's apartment building rather than the front and then used StreetViewAI to virtually walk around the building.

7.4 Overarching Themes

At the end of the study, participants were asked debrief questions, including *"How well did StreetViewAI support you in finding the information you needed?"*, *"How valuable was the information StreetViewAI provided?"*, and *"How accurate did you perceive the information?"*—see Table 8. In general, participants felt that the tool supported them in finding information and that the information itself was valuable but had some concerns about accuracy. Below, we report on cross-cutting themes extending the sections above.

	POI Investigation			Navigation		
	Mean	Median	SD	Mean	Median	SD
How well	6.0	6.0	0.7	6.0	6.0	0.9
How valuable	6.4	7.0	0.6	6.1	6.5	1.1
How accurate	5.7	6.0	0.9	5.5	5.5	1.2

Table 8: *Post-hoc* 7-point Likert scale ratings (7 is best).

Perceptions of accuracy. Interestingly, participants generally rated StreetViewAI's accuracy as high (*Med*=6.0 for POI investigation and 5.5 for navigation) even though, as P1 said, *"It seems pretty accurate but I don't have vision so I don't really know"* and P10 said *"It didn't come up with hallucinatory answers—as far as I know"*. These perceptions were likely influenced by the intermixing of two data sources, geographic data from Google Maps (which is highly accurate) and AI descriptions (which may not be). For example, P11 experienced a problem in the POI task where the AI chat said that there was no Mexican restaurant nearby, however, as they rotated their view, they heard the restaurant reported in the list of local nearby places close to them (which is not inferenced). *"She [the AI] got confused about the Mexican restaurant. But when I was spinning around, she did a great job describing what was around."* Similarly, P8 felt that the tool was hallucinating but then used hotkeys to confirm information about the current road and nearby places.

Conceptualizing street view and controls. As we saw with our co-design, it is challenging to explain the concept of high-resolution, immersive 360° streetscape imagery and navigating therein to blind users. After the third participant, we changed our study introduction to incorporate role play: participants physically stood up and rotated their bodies 45° to the right (to demonstrate →) and 45° to the left (for ←). Participants also took a physical step along their current heading and compared this to virtual steps in StreetViewAI (which were 10-15 meters). This strengthened understanding and led to fewer confusions, especially for initial tasks. We return to disembodied control in the Discussion.

Building spatial awareness. As evident by the number of orientation-related AI chats, participants focused heavily on building a mental model of a space—where the person was inside the virtual-physical world, where objects were, and there relative positions. Though imperfect, the MLLM surprised us in some cases; for example, in imagining new orientations and providing guidance:

> ***P9:*** *Let's say I stepped out of the restaurant, can you orient me from that perspective?*
> ***AI:*** *Okay, if you were to step out of the restaurant, you would be facing the street. You would have the Thompson St. sign on your left and be facing Southeast.*

The interactive chat allowed participants to orient their perspectives and strengthen their mental models:

P10: *Where is the door?*
AI: *The door to the cafe is directly in front of you.*
P10: *Do I have to cross the street to get to it?*
AI: *No, you do not have to cross the street to get to the door. It is on your side of the street.*
P10: *Is it on my left or right side?*
AI: *The door is directly in front of you, not to your left or right.*
P10: *That's interesting! I thought the intersection was in front of me.*
AI: *You're right, the intersection is also straight ahead, and the door is just to the left of that. So, while the intersection is straight ahead, the door is slightly to your left relative to the intersection.*

Still, because StreetViewAI intermixes both relative (*"On your left..."*) and absolute positions (*"You're facing northeast"*), some participants struggled to make sense of this information, especially since panning status messages always used absolute references.

Alternative access to information. Finally, we asked participants *"If you didn't have a tool like StreetViewAI, how would you otherwise find out the information it provides"*. P3 said, starkly, *"I wouldn't."* but most participants mentioned relying on others: *"I would bring someone with me [to the location] and ask them all the questions"* (P4), *"Using paratransit and asking people"* (P6), or *"Asking people on AIRA [3]—[a remote video interpreting service]"*. P7, P8, and P10 mentioned existing accessible *in situ* navigation tools like Audio Guidance in Google Maps [41] or VoiceVista [73] and walking around physically. In comparison to those tools, however, P10 emphasized: *"What I like about this [StreetViewAI], is that I can ask questions and be proactive—that's a real game changer for me."*

8 Discussion

We introduced and evaluated *StreetViewAI* (Figure 1), the first accessible street view tool for BLV users using context-aware AI and keyboard-based navigation controls. Our findings help highlight *what* information BLV users desire from and ask about streetscape imagery, the potential of multimodal AI models to appropriately answer contextualized inquiries, and uncovers important implications for the design of future accessible streetscape tools. We reflect on key findings, limitations, and opportunities for future work.

8.1 Interacting with AI-Powered Streetscapes

Our study revealed valuable insights into how blind users engage with and perceive an AI-driven streetscape environment. A key finding was the participants' preference for the AI Chat Agent over passively receiving triggered descriptions (AI Describer via Alt + D), suggesting a desire for agency and targeted information seeking. Unlike AI Describer, the chat system is intrinsically personalized to questions of interest. However, the reliance on AI for sensemaking and navigating complex visual scenes also surfaced potential issues around accuracy and trust, data sources and discrepancies, and the ability for the AI backend to go beyond the user's current context to address questions.

Accuracy and trust. Participants exhibited a high-level of trust in StreetViewAI, even with inaccuracies. As other scholars have emphasized [75, 97, 131], it is difficult to discern hallucinations from true information—StreetViewAI seems just as confident about both. For example, StreetViewAI repeatedly told participants that there was no bike rack at the Mexican restaurant but there was (though in some sessions, it got it right). StreetViewAI's ability to respond and provide human-like answers for most questions—whether they were correct—increased perceptions of trust and accuracy: *"There are so few things that the system seemed like it could not answer"* (P10). While the risks in a virtual environment are lower here than relying on real-time AI guidance in the physical world, the potentially flawed mental models of geography and spatial relationships could still have profound effects on planning travel and safety. Future work should explore how best to frame information gained from StreetViewAI and limitations therein, borrowing from literature in Explainable AI (XAI) [14].

Discrepancies in data. StreetViewAI draws primarily on two information sources: geographic databases of road, place, and address information (which is up-to-date and trustworthy) and AI-generated descriptions about the scene and local geography (which can depend on outdated street view imagery and imperfect inferences). Occasionally, these data sources would disagree: *e.g.*,, for one participant, the AI said that there was *"no Mexican restaurant in front of them"* even though the geographic-based status message said that there was. Interestingly, when specifically asked about whether participants *thought* about streetscape image age when using StreetViewAI in the post-study debrief, only a few did—despite being taught how to use Alt + P to hear metadata about the GSV photograph, including capture date.

Increasing data sources. Over 70% of non-responses from the AI (40/54) were due to a lack of connected data sources, including transit schedules, nearby places beyond the immediate area (> 50 meter context radius), and restaurant menus. These queries should all be possible in the future by feeding additional data into the prompts or by connecting the multimodal AI to external knowledge sources (*e.g.,* via function calls). Similarly, to address limitations in streetscape image age, future work should incorporate additional image datasets (*e.g.,* the Google Places API contains millions of business- and user-contributed photos [57], which are often up-to-date and show both indoor and outdoor imagery).

Towards a more autonomous AI agent. While the AI Chat Agent had access to all previous user views—from heading changes and movement—in its context window (up to ~1M input tokens), StreetViewAI could not address queries that went beyond these views unless that data was provided as geographic context (*e.g.,* nearby places within 50 meters, nearby roads, *etc.*). Future work should create a more autonomous agentic backend capable of examining the full 360° at every location and all nearby panos and metadata to address user queries. For example, *"What's the next bus stop down this road and what does it look like?"*. Here, the AI agent should automatically query geographic knowledge sources, find the bus stop location, capture and analyze surrounding GSV imagery, and return a response. The user could then inquire about the walking route to the bus stop, which the AI agent could again automatically analyze using both geographic metadata as well as CV analysis of GSV images along the way.

8.2 Navigation, Orientation, and Routing

Effectively navigating the virtual environment itself presented distinct challenges and opportunities for our blind participants, including user orientation, precise placement within the scene, and desired pathfinding capabilities.

Orientation. In a virtual, disembodied environment, supporting and maintaining the user's understanding of their orientation with respect to the current location, view, and spatial relationships therein is a significant challenge. We found that standing and physically role playing turns *vs.* steps helped develop appropriate mental models of the movement controls but that maintaining spatial orientation was difficult. Some users strongly preferred relative directions only (*e.g., "On my left..."* or *"In front of me..."* despite absolute directions (cardinal and intercardinal headings) having clear advantages of being non-contextual. These challenges may fade as users gain more experience.

The importance of landing. When teleporting to an address or POI, we used a custom algorithm to automatically turn the user's view towards that destination; however, this technique was imperfect and the cost of being wrong and recovery time for a blind user *vs.* a sighted user is disproportionately high. Even if the heading angle was off by a small amount (~45°) or a sub-optimal pano was selected (*e.g.,* 10 meters from the destination), the blind user needs to first *discover* the error (*e.g., "I'm not looking at what I thought..."*) and then recover (*e.g., "Which direction should I be facing? Where is the destination from here?"*). Future work should use CV analysis of the surrounding drop points and determine the best pano and view for a blind user to investigate the destination (*e.g.,* a store's entrance and the pedestrian pathway in front), or could ask the user which of several options they prefer.

Routing. Building on the above, while StreetViewAI currently supports POI investigations, open-world navigation, and virtual tourism, it does not yet support origin-to-destination routing—a key task identified in our co-design sessions. In contrast to open-world navigation where StreetViewAI knows the origin but not the destination, in routing, the tool would know both. At a minimum, supporting this feature would enable users to virtually travel a route with turn-by-turn navigation (similar to *Voice Guidance* in Google Maps [41]). As P9 said, *"I want to ask about the 'last mile' walk from the nearest bus stop to the cafe"*. Other interesting features are also worth exploring, such as: (1) using an AI agent to examine all panos along potential routes and determining potential obstacles; (2) creating blind-friendly summaries of the route; and (3) helping with the "last 10 meters" [105], including door finding.

Bookmarking. Finally, similar to *in situ* navigation tools such as *VoiceVista* [73] and *FootNotes* [38], users wanted the ability to bookmark locations, annotate them, and then easily retrieve them either again on the laptop or while actually physically navigating in the real world on their smartphone.

8.3 Audio Descriptions and Beyond

In StreetViewAI, the primary output modality is discrete verbalized messages. We reflect on content, concision, and other possibilities.

Ephemeral data. In the post-study debrief, participants emphasized that certain vocalized details were irrelevant and distracting, such as ephemeral objects (*"That parked car would not still be there though, right?"*), car colors, or the presence of pedestrians or traffic. While sighted users of GSV can easily ignore these visual details—and, indeed, may not even notice them—they take up valuable audio bandwidth if described by the AI model. Still, some participants felt that, with caveats, such information was useful to determine, for example, how busy a location could be. Future tools should enable users to customize what features are most important to describe and what to ignore.

Concise announcements. To maintain user orientation and support their mental model of the local geography and navigable areas, we provide status updates at every interaction from heading changes to movement; however, constructing these messages for clarity, relevance, and concision is challenging. In initial prototypes, we triggered AI Describer descriptions at every heading change or movement. In later prototypes, including that used in our study, we decided to require explicit hot key presses (Alt + D). Still, these status updates could be too verbose, which would cause frustration and lead to information loss (as important messages are diluted). Participants suggested being able to customize the output for their current tasks both in terms of what the AI describes as well as what is included in the status updates.

Going beyond audio descriptions. As an initial prototype, we relied on discrete audio descriptions for both navigational feedback and AI descriptions (earcons were used to notify the user about microphone toggling). However, many accessible navigation tools [90, 93, 118], VR techniques [88, 108, 110, 120], and audio games [10, 101] use spatialized audio or haptics to provide nonverbal cues about distances, object locations, and other information. Future work should examine these additional, complementary output modalities (*e.g.,* modulating a musical pitch for heading orientation). Unlike digital content in VR or gaming—where exact distances can be trivially calculated—it is more challenging in streetscape images; however, new monocular depth estimation algorithms from RGB images [119, 129] could be used to auto-generate 3D audio soundscapes. Moreover, the sounds of the streetscape could be synthesized and spatially rendered based on emerging image-to-audio synthesizes [112, 140] (*e.g.,* hearing the sounds of a cafe or street traffic when turning towards it).

Geo-spatial voice agents for all. While our primary goal was to create an accessible streetscape tool, as development progressed and we added bidirectional voice conversation capabilities and voice commands, we realized that StreetViewAI could be the beginning of a generalizable audio-only streetscape experience. Imagine, for example, conversing with a voice agent like *Alexa* or *Siri* while driving and learning about the color of a building at an upcoming turn or whether there appears to be a nearby parking lot and where. Such queries could then trigger an AI-based streetscape analysis along the user's route or destination.

8.4 Limitations

Our study had three primary limitations. First, while we recruited eleven blind participants, none were familiar with Mac keyboards and some preferred smartphones as their primary computing device; thus, the dominant use of voice *vs.* typing for AI Chat (> 90%) may be artificially inflated. As P10 said: "*If I were more comfortable with your keyboard, I would use* Alt+D *to get a Gestalt first* *and then* Alt+C *to drill down more*." Moreover, while our participants varied in age and demographics, all were cane users and most were technology savvy with medium-to-high AI familiarity. Second, while our question type analysis reveals important findings about *what* information participants are interested in deriving from streetscape imagery—including spatial orientation, the existence or presence of objects, describing the current view—the frequency and nature of questions asked are a reflection of the study tasks as well as our post-task questions. Derived from literature and our co-design sessions, we believe the study tasks had high ecological validity; however, future work should explore real usage of StreetViewAI in a deployment study. Third, and finally, due to study length constraints, we could not examine all parts of the StreetViewAI prototype, including the *AI Tour Guide* and setting a custom user profile. However, these features were co-designed and tested by our blind collaborators; future work should study AI personas and user profiles in more detail.

9 Conclusion

StreetViewAI marks a fundamental advance in making immersive streetscape environments accessible. As P8 said, *"Google's Street View tool is not accessible at all... this is a huge leap forward in navigation"*. Our work contributes not only a novel system but also design insights for interacting with 360° imagery non-visually, interaction techniques for conversational AI agents in spatial contexts, and a deeper understanding of the information needs of BLV users exploring virtual representations of the physical world.

Acknowledgments

Diagram icons from Noun Project, including: "prompt icon" by Firdaus Faiz, "command functions" by Kawalan Icon, "<lat,lng>" by Didik Darmanto, "heading" and "relative heading" by IronCV, "road address" and "place address" by IGraphics, "city, state, country" by Karyative, "café" by loviana, "hospital" by loviana, "park" by Made x Made, "school" by Omah Icon, "storefront" by Sutiyah, "place type" by BEARicons, "editorial summary" by ilham firmansyah, "relative distance" by Sandy Walsh, "blind person" by Nuno Sequeira, "game controller" by PMSO BEMFEBUNUD, "up arrow key" by Se-hui Jo, "left arrow key" by Se-hui Jo, "MLLM icon" by Funtasticon.

References

[1] Dragan Ahmetovic, Federico Avanzini, Adriano Baratè, Cristian Bernareggi, Gabriele Galimberti, Luca A. Ludovico, Sergio Mascetti, and Giorgio Presti. 2019. Sonification of Rotation Instructions to Support Navigation of People with Visual Impairment. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom.* 1–10. doi:10.1109/PERCOM.2019.8767407

[2] Dragan Ahmetovic, Cole Gleason, Chengxiong Ruan, Kris Kitani, Hironobu Takagi, and Chieko Asakawa. 2016. NavCog: a navigational cognitive assistant for the blind. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 90–99. doi:10.1145/2935334.2935361

[3] Aira. 2025. Aira: Video Remote Interpreting. https://aira.io/ Remote video interpreting service for blind, low-vision, deaf, and hard-of-hearing communities.

[4] Temitope Akinboyewa, Zhenlong Li, Huan Ning, and M. Naser Lessani. 2024. GIS Copilot: Towards an Autonomous GIS Agent for Spatial Analysis. arXiv:2411.03205 [cs.AI] https://arxiv.org/abs/2411.03205

[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* (Melbourne, Australia) *(OzCHI '18)*. Association for Computing Machinery, New York, NY, USA, 278–289. doi:10.1145/3292147.3292163

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[7] Apple Inc. 2024. Get live descriptions of your surroundings on iPhone. Apple Support. https://support.apple.com/guide/iphone/get-live-descriptions-of-your-surroundings-iph37e6b3844/ios Accessed: April 9, 2025.

[8] Apple Inc. 2024. Use VoiceOver Recognition on your iPhone or iPad. Apple Support. https://support.apple.com/en-us/111799 Accessed: April 9, 2025.

[9] Apple Inc. 2025. Apple Look Around. https://www.apple.com/maps/

[10] Matthew Tylee Atkinson. 2018. AGRIP: Accessible Gaming Rendering Independence Possible / AudioQuake. https://github.com/matatk/agrip

[11] Matthew T. Atkinson, Sabahattin Gucukoglu, Colin H. C. Machin, and Adrian E. Lawrence. 2006. Making the mainstream accessible: redefining the game. In *Proceedings of the 2006 ACM SIGGRAPH Symposium on Videogames* (Boston, Massachusetts) *(Sandbox '06)*. Association for Computing Machinery, New York, NY, USA, 21–28. doi:10.1145/1183316.1183321

[12] Vahid Balali, Armin Ashouri Rad, and Mani Golparvar-Fard. 2015. Detection, classification, and mapping of U.S. traffic signs using google street view images for roadway inventory management. *Visualization in Engineering* 3, 1 (11 2015), 15. doi:10.1186/s40327-015-0027-1

[13] Maryam Bandukda, Catherine Holloway, Aneesha Singh, Giulia Barbareschi, and Nadia Berthouze. 2021. Opportunities for Supporting Self-efficacy Through Orientation & Mobility Training Technologies for Blind and Partially Sighted People. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 32, 13 pages. doi:10.1145/3441852.3471224

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[15] Be My Eyes. 2025. Be My AI. https://www.bemyeyes.com/blog/be-my-ai-launches-globally

[16] Anke M. Brock, Jon E. Froehlich, João Guerreiro, Benjamin Tannert, Anat Caspi, Johannes Schöning, and Steve Landau. 2018. SIG: Making Maps Accessible and Putting Accessibility in Maps. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3170427.3185373

[17] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 140, 18 pages. doi:10.1145/3654777.3676375

[18] 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. arXiv:2502.03564 [cs.HC] https://arxiv.org/abs/2502.03564

[19] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2018), 834–848. doi:10.1109/TPAMI.2017.2699184

[20] Arnavi Chheda-Kothary, Ritesh Kanchi, Chris Sanders, Kevin Xiao, Aditya Sengupta, Melanie Kneitmix, Jacob O. Wobbrock, and Jon E. Froehlich. 2025. ArtInsight: Enabling AI-Powered Artwork Engagement for Mixed Visual-Ability Families. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 190–210. doi:10.1145/3708359.3712082

[21] Arnavi Chheda-Kothary, Jacob O. Wobbrock, and Jon E. Froehlich. 2024. Engaging with Children's Artwork in Mixed Visual-Ability Families. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 3, 19 pages. doi:10.1145/3663548.3675613

[22] Victoria Clarke and Virginia Braun and. 2017. Thematic analysis. *The Journal of Positive Psychology* 12, 3 (2017), 297–298. doi:10.1080/17439760.2016.1262613 arXiv:https://doi.org/10.1080/17439760.2016.1262613

[23] Jazmin Collins, Kaylah Myranda Nicholson, Yusuf Khadir, Andrea Stevenson Won, and Shiri Azenkot. 2024. An AI Guide to Enhance Accessibility of Social Virtual Reality for Blind People. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 130, 5 pages. doi:10.1145/3663548.3688498

[24] Megan Conway, Brett Oppegaard, and Tuyet Hayes. 2020. Audio description: Making useful maps for blind and visually impaired people. *Technical Communication* 67, 2 (2020), 68–86.

[25] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Kirk Andrew Crawford, Jennifer Posada, Yetunde Esther Okueso, Erin Higgins, Laura Lachin, and Foad Hamidi. 2024. Co-designing a 3D-Printed Tactile Campus Map With Blind and Low-Vision University Students. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 77, 6 pages. doi:10.1145/3663548.3688537

[27] Yuri Danilov and Mitchell Tyler. 2005. Brainport: An Alternative Input to the Brain. *Journal of Integrative Neuroscience* 04, 04 (2005), 537–550. doi:10.1142/S0219635205000914 arXiv:https://doi.org/10.1142/S0219635205000914

[28] Franco Delogu, Massimiliano Palmiero, Stefano Federici, Catherine Plaisant, Haixia Zhao, and Olivetti Belardinelli and. 2010. Non-visual exploration of geographic maps: Does sonification help? *Disability and Rehabilitation: Assistive Technology* 5, 3 (2010), 164–174. doi:10.3109/17483100903100277 arXiv:https://doi.org/10.3109/17483100903100277 PMID: 20415530.

[29] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2022. Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. *Journal of Artificial Intelligence Research* 73 (Jan. 2022), 437–459. doi:10.1613/jair.1.13113

[30] Michael Duan, Shosuke Kiami, Logan Milandin, Johnson Kuang, Michael Saugstad, Maryam Hosseini, and Jon E. Froehlich. 2022. Scaling Crowd+AI Sidewalk Accessibility Assessments: Initial Experiments Examining Label Quality and Cross-city Training on Performance. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 82, 5 pages. doi:10.1145/3517428.3550381

[31] Julie Ducasse, Anke M. Brock, and Christophe Jouffrais. 2018. *Accessible Interactive Maps for Visually Impaired Users.* Springer International Publishing, Cham, 537–584. doi:10.1007/978-3-319-54446-5_17

[32] John Dudley, Lulu Yin, Vanja Garaj, and Per Ola Kristensson. 2023. Inclusive Immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse. *Virtual Reality* 27, 4 (12 2023), 2989–3020. doi:10.1007/s10055-023-00850-8

[33] Freedom Scientific. 2025. JAWS Screen Reader. https://www.freedomscientific.com/products/software/jaws/

[34] Jon E. Froehlich, Anke M. Brock, Anat Caspi, João Guerreiro, Kotaro Hara, Reuben Kirkham, Johannes Schöning, and Benjamin Tannert. 2019. Grand challenges in accessible maps. 26, 2 (Feb. 2019), 78–81. doi:10.1145/3301657

[35] GMA Games. 2025. Shades of Doom. https://www.gmagames.com/sod.html

[36] Nicholas A. Giudice and Gordon E. Legge. 2008. Blind Navigation and the Role of Technology. In *Engineering Handbook of Smart Technology for Aging, Disability, and Independence*, Abdelsalam Helal, Mounir Mokhtari, and Bessam Abdulrazak (Eds.). John Wiley & Sons, 479–500. https://umaine.edu/vemi/wp-content/uploads/sites/220/2016/08/GiudiceLegge-2008-BLIND-NAVIGATION-AND-THE-ROLE-OF-TECHNOLOGY-EHST.pdf

[37] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 549–559. doi:10.1145/3308558.3313605

[38] Cole Gleason, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2018. FootNotes: Geo-referenced Audio Annotations for Nonvisual Exploration. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 109 (Sept. 2018), 24 pages. doi:10.1145/3264919

[39] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 901, 21 pages. doi:10.1145/3613904.3642211

[40] GoodMaps Inc. 2025. GoodMaps. https://goodmaps.com/

[41] Google. 2023. Voice Guidance in Maps, Built for People with Impaired Vision. https://blog.google/products/maps/better-maps-for-people-with-vision-impairments/ Official Google blog post about accessibility features in Google Maps.

[42] Google. 2024. Use image descriptions in TalkBack. Android Accessibility Help. https://support.google.com/accessibility/android/answer/15341968 Accessed: April 9, 2025.

[43] Google. 2025. Celebrate 15 years of exploring your world on Street View. https://www.google.com/streetview/anniversary/

[44] Google. 2025. Google Cloud Speech-to-Text API. https://cloud.google.com/speech-to-text

[45] Google. 2025. Google Cloud Text-to-Speech API. https://cloud.google.com/text-to-speech

[46] Google. 2025. Google Gemini API. https://ai.google.dev/gemini-api/docs/models

[47] Google. 2025. Google Maps Places API. https://developers.google.com/maps/documentation/places

[48] Google. 2025. Google Maps Roads API. https://developers.google.com/maps/documentation/roads

[49] Google. 2025. Google Maps Street View API. https://developers.google.com/maps/documentation/streetview

[50] Google. 2025. Google Street View. https://www.google.com/streetview/

[51] Google. 2025. Google Vertex AI. https://cloud.google.com/vertex-ai

[52] Google. 2025. Supported voices and languages - Google Cloud Text-to-Speech. https://cloud.google.com/text-to-speech/docs/list-voices-and-types#list_of_all_supported_languages

[53] Google. 2025. Treks: Grand Canyon. https://www.google.com/maps/about/behind-the-scenes/streetview/treks/grand-canyon/

[54] Google. 2025. *Vertex AI Multimodal Live API.* Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal-live-api

[55] Google DeepMind. 2024. Project Astra. https://deepmind.google/technologies/project-astra/. Accessed: 2025-04-08.

[56] Google DeepMind. 2025. Gemini model updates: February 2025. https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/

[57] Google LLC. 2024. Google Places API: The Place Photo Service. Google Developers Documentation. https://developers.google.com/maps/documentation/places/web-service/photos Accessed: April 9, 2025.

[58] Christina Granquist, Susan Y. Sun, Sandra R. Montezuma, Tu M. Tran, Rachel Gage, and Gordon E. Legge. 2021. Evaluation and Comparison of Artificial Intelligence Vision Aids: Orcam MyEye 1 and Seeing AI. *Journal of Visual Impairment & Blindness* 115, 4 (2021), 277–285. doi:10.1177/0145482X211027492 arXiv:https://doi.org/10.1177/0145482X211027492

[59] Ananya Gubbi Mohanbabu and Amy Pavel. 2024. Context-Aware Image Descriptions for Web Accessibility. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 62, 17 pages. doi:10.1145/3663548.3675658

[60] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. 2013. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with Google street view. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 16, 8 pages. doi:10.1145/2513383.2513448

[61] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. 2015. Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View: An Extended Analysis. *ACM Trans. Access. Comput.* 6, 2, Article 5 (March 2015), 23 pages. doi:10.1145/2717513

[62] Kotaro Hara, Jin Sun, Noa Chazan, David Jacobs, and Jon E. Froehlich. 2013. An Initial Study of Automatic Curb Ramp Detection with Crowdsourced Verification Using Google Street View Images. In *Extended Abstracts of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP-2013)*. HCOMP-2013.

[63] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. 2014. Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 189–204. doi:10.1145/2642918.2647403

[64] Wilko Heuten, Daniel Wichmann, and Susanne Boll. 2006. Interactive 3D sonification for the exploration of city maps. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles* (Oslo, Norway) *(NordiCHI '06)*. Association for Computing Machinery, New York, NY, USA, 155–164. doi:10.1145/1182475.1182492

[65] Leona Holloway, Matthew Butler, and Kim Marriott. 2023. TactIcons: Designing 3D Printed Map Icons for People who are Blind or have Low Vision. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 543, 18 pages. doi:10.1145/3544548.3581359

[66] Leona Holloway, Kim Marriott, and Matthew Butler. 2018. Accessible Maps for the Blind: Comparing 3D Printed Models with Tactile Graphics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173772

[67] Leona Holloway, Kim Marriott, Matthew Butler, and Samuel Reinders. 2019. 3D Printed Maps and Icons for Inclusion: Testing in the Wild by People who are Blind or have Low Vision. In *Proceedings of the 21st International ACM*

SIGACCESS Conference on Computers and Accessibility (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 183–195. doi:10.1145/3308561.3353790

[68] Jonggi Hong, Alisha Pradhan, Jon E. Froehlich, and Leah Findlater. 2017. Evaluating Wrist-Based Haptic Feedback for Non-Visual Target Finding and Path Tracing on a 2D Surface. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 210–219. doi:10.1145/3132525.3132538

[69] Maryam Hosseini, Fabio Miranda, Jianzhe Lin, and Claudio T. Silva. 2022. CitySurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society* 79 (2022), 103630. doi:10.1016/j.scs.2021.103630

[70] Maryam Hosseini, Mikey Saugstad, Fabio Miranda, Andres Sevtsuk, Claudio T. Silva, and Jon E. Froehlich. 2022. Towards Global-Scale Crowd+AI Techniques to Map and Assess Sidewalks for People with Disabilities. arXiv:2206.13677 [cs.CV] https://arxiv.org/abs/2206.13677

[71] Mihail Ivanchev, Francis Zinke, and Ulrike Lucke. 2014. Pre-journey Visualization of Travel Routes for the Blind on Refreshable Interactive Tactile Displays. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Deborah Fels, Dominique Archambault, Petr Peňáz, and Wolfgang Zagler (Eds.). Springer International Publishing, Cham, 81–88.

[72] Junehyung Jeon and Ayoung Woo. 2023. Deep learning analysis of street panorama images to evaluate the streetscape walkability of neighborhoods for subsidized families in Seoul, Korea. *Landscape and Urban Planning* 230 (2023), 104631. doi:10.1016/j.landurbplan.2022.104631

[73] Jianfeng Wu. 2025. VoiceVista. https://drwjf.github.io/vvt/

[74] Lucy Jiang, Jon E. Froehlich, and Leah Findlater. 2024. Making Urban Art Accessible: Current Art Access Techniques, Design Considerations, and the Role of AI. In *The 4th Annual Workshop on The Future of Urban Accessibility: The Role of AI* (Virtual). 8 pages.

[75] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Trang Thi Thu Xuan, Essi Häyhänen, Jinan Y. Azem, and Bernard J. Jansen. 2025. "You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1624–1638. doi:10.1145/3708359.3712160

[76] Vaishnav Kameswaran, Alexander J. Fiannaca, Melanie Kneisel, Amy Karlson, Edward Cutrell, and Meredith Ringel Morris. 2020. Understanding In-Situ Use of Commonly Available Navigation Technologies by People with Visual Impairments. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 28, 12 pages. doi:10.1145/3373625.3416995

[77] KRDO. 2022. Google's new Street View camera is more portable and looks weirdly cute. https://krdo.com/news/2022/05/24/googles-new-street-view-camera-is-more-portable-and-looks-weirdly-cute-2/

[78] Minchu Kulkarni, Chu Li, Jaye Jungmin Ahn, Katrina Oi Yau Ma, Zhihan Zhang, Michael Saugstad, Kevin Wu, Yochai Eisenberg, Valerie Novack, Brent Chamberlain, and Jon E. Froehlich. 2023. BusStopCV: A Real-time AI Assistant for Labeling Bus Stop Accessibility Features in Streetscape Imagery. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 91, 6 pages. doi:10.1145/3597638.3614481

[79] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 462, 15 pages. doi:10.1145/3491102.3501966

[80] Jiyun Lee, Donghyun Kim, and Jina Park. 2022. A Machine Learning and Computer Vision Study of the Environmental Characteristics of Streetscapes That Affect Pedestrian Satisfaction. *Sustainability* 14, 9 (2022). doi:10.3390/su14095730

[81] Jaewook Lee, Yang Li, Dylan Bunarto, Eujean Lee, Olivia H. Wang, Adrian Rodriguez, Yuhang Zhao, Yapeng Tian, and Jon E. Froehlich. 2024. Towards AI-Powered AR for Enhancing Sports Playability for People with Low Vision: An Exploration of ARSports. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 228–233. doi:10.1109/ISMAR-Adjunct64951.2024.00055

[82] Jaewook Lee, Andrew D. Tjahjadi, Jiho Kim, Junpu Yu, Minji Park, Jiawen Zhang, Jon E. Froehlich, Yapeng Tian, and Yuhang Zhao. 2024. CookAR: Affordance Augmentations in Wearable AR to Support Kitchen Tool Interactions for People with Low Vision. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 141, 16 pages. doi:10.1145/3654777.3676449

[83] Chu Li, Rock Yuren Pang, Ather Sharif, Arnavi Chheda-Kothary, Jeffrey Heer, and Jon E. Froehlich. 2024. AltGeoViz: Facilitating Accessible Geovisualization. In *2024 IEEE Visualization and Visual Analytics (VIS)*. 61–65. doi:10.1109/VIS55277.2024.00020

[84] Lit. 2025. Lit: Simple. Fast. Web Components. https://github.com/lit/lit

[85] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[86] Ali Mansourian and Rachid Oucheikh. 2024. ChatGeoAI: Enabling Geospatial Analysis for Public through Natural Language, with Large Language Models. *ISPRS International Journal of Geo-Information* 13, 10 (2024). doi:10.3390/ijgi13100348

[87] Mapillary. 2025. About Mapillary. https://www.mapillary.com/about

[88] Keenan R. May, Brianna J. Tomlinson, Xiaomeng Ma, Phillip Roberts, and Bruce N. Walker. 2020. Spotlights and Soundscapes: On the Design of Mixed Reality Auditory Environments for Persons with Visual Impairment. *ACM Trans. Access. Comput.* 13, 2, Article 8 (April 2020), 47 pages. doi:10.1145/3378576

[89] Meta Platforms, Inc. 2025. Mapillary by Meta. https://www.mapillary.com/

[90] Microsoft. 2023. *Microsoft Soundscape: Empowering people who are blind or have low vision to explore the world around them.* Microsoft. https://blogs.microsoft.com/accessibility/soundscape/ Microsoft Accessibility Blog.

[91] Microsoft Corporation. 2025. Bing Streetside. https://www.bing.com/maps/streetside

[92] Microsoft Corporation. 2025. Seeing AI. https://www.microsoft.com/en-us/ai/seeing-ai

[93] MIPsoft. 2025. BlindSquare. https://www.blindsquare.com/

[94] MobX. 2025. MobX: Simple, scalable state management. https://github.com/mobxjs/mobx

[95] 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 Computer Society, Los Alamitos, CA, USA, 451–454. doi:10.1109/ISMAR-Adjunct.2019.00122

[96] Ruth G Nagassa, Matthew Butler, Leona Holloway, Cagatay Goncu, and Kim Marriott. 2023. 3D Building Plans: Supporting Navigation by People who are Blind or have Low Vision in Multi-Storey Buildings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 539, 19 pages. doi:10.1145/3544548.3581389

[97] Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations. arXiv:2404.03745 [cs.HC] https://arxiv.org/abs/2404.03745

[98] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[99] Konstantinos Papadopoulos, Eleni Koustriava, Panagiotis Koukourikos, Lefkothea Kartasidou, Marialena Barouti, Asimis Varveris, Marina Misiou, Timoclia Zacharogeorga, and Theocharis Anastasiadis and. 2017. Comparison of three orientation and mobility aids for individuals with blindness: Verbal description, audio-tactile map and audio-haptic map. *Assistive Technology* 29, 1 (2017), 1–7. doi:10.1080/10400435.2016.1171809 arXiv:https://doi.org/10.1080/10400435.2016.1171809 PMID: 27184731.

[100] Maria Teresa Paratore and Barbara Leporini. 2024. Exploiting the haptic and audio channels to improve orientation and mobility apps for the visually impaired. *Universal Access in the Information Society* 23, 2 (6 2024), 859–869. doi:10.1007/s10209-023-00973-4

[101] Jamie Pauls. 2020. Vintage Games Series, Part 4: Immerse Yourself in the World of Shades of Doom. https://www.afb.org/aw/21/12/17336

[102] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs.CV] https://arxiv.org/abs/1506.02640

[103] Samuel Reinders, Matthew Butler, Ingrid Zukerman, Bongshin Lee, Lizhen Qu, and Kim Marriott. 2025. When Refreshable Tactile Displays Meet Conversational Agents: Investigating Accessible Data Presentation and Analysis with Touch and Speech. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 864–874. doi:10.1109/TVCG.2024.3456358

[104] RxJS. 2025. RxJS: Reactive Extensions Library for JavaScript. https://github.com/ReactiveX/rxjs

[105] Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 222–235. doi:10.1145/3308561.3353776

[106] Daisuke Sato, Uran Oh, João Guerreiro, Dragan Ahmetovic, Kakuya Naito, Hironobu Takagi, Kris M. Kitani, and Chieko Asakawa. 2019. NavCog3 in the Wild: Large-scale Blind Indoor Navigation Assistant with Semantic Features. *ACM Trans. Access. Comput.* 12, 3, Article 14 (Aug. 2019), 30 pages. doi:10.1145/3340319

[107] Joram Schito and Sara Irina Fabrikant and. 2018. Exploring maps by sounds: using parameter mapping sonification to make digital elevation models audible. *International Journal of Geographical Information Science* 32, 5 (2018), 874–906. doi:10.1080/13658816.2017.1420192 arXiv:https://doi.org/10.1080/13658816.2017.1420192

[108] David W. Schloerb, Orly Lahav, Joseph G. Desloge, and Mandayam A. Srinivasan. 2010. BlindAid: Virtual environment system for self-reliant trip planning and orientation and mobility training. In *2010 IEEE Haptics Symposium*. 363–370. doi:10.1109/HAPTIC.2010.5444631

[109] Ather Sharif, Andrew Mingwei Zhang, Anna Shih, Jacob O. Wobbrock, and Katharina Reinecke. 2022. Understanding and Improving Information Extraction From Online Geospatial Data Visualizations for Screen-Reader Users. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 61, 5 pages. doi:10.1145/3517428.3550363

[110] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376353

[111] Maximilian Speicher, Brian D. Hall, and Michael Nebeling. 2019. What is Mixed Reality?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300767

[112] Xia Su, Jon E. Froehlich, Eunyee Koh, and Chang Xiao. 2024. SonifyAR: Context-Aware Sound Generation in Augmented Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 128, 13 pages. doi:10.1145/3654777.3676406

[113] Sarit Szpiro, Yuhang Zhao, and Shiri Azenkot. 2016. Finding a store, searching for a product: a study of daily challenges of low vision people. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 61–72. doi:10.1145/2971648.2971723

[114] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, and Andrew M. Dai et al. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[115] Kavita E. Thomas, Livia Sumegi, Leo Ferres, and Somayajulu Sripada. 2008. Enabling access to geo-referenced information: Atlas.txt. In *Proceedings of the 2008 International Cross-Disciplinary Conference on Web Accessibility (W4A)* (Beijing, China) *(W4A '08)*. Association for Computing Machinery, New York, NY, USA, 101–104. doi:10.1145/1368044.1368066

[116] Bertrand Tornil and Nadine Baptiste-Jessel. 2004. Use of Force Feedback Pointing Devices for Blind Users. In *User-Centered Interaction Paradigms for Universal Access in the Information Society*, Christian Stary and Constantine Stephanidis (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 479–485.

[117] Bruce N. Walker and Jeffrey Lindsay and. 2005. Using Virtual Environments to Prototype Auditory Navigation Displays. *Assistive Technology* 17, 1 (2005), 72–81. doi:10.1080/10400435.2005.10132097 arXiv:https://doi.org/10.1080/10400435.2005.10132097 PMID: 16121647.

[118] Bruce N. Walker and Jeffrey Lindsay. 2006. Navigation Performance With a Virtual Auditory Display: Effects of Beacon Sound, Capture Radius, and Practice. *Human Factors* 48, 2 (2006), 265–278. doi:10.1518/001872006777724507 arXiv:https://doi.org/10.1518/001872006777724507 PMID: 16884048.

[119] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. 2025. CRM: Single Image to 3D Textured Mesh with Convolutional Reconstruction Model. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 57–74.

[120] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300371

[121] Galen Weld, Esther Jang, Anthony Li, Aileen Zeng, Kurtis Heimerl, and Jon E. Froehlich. 2019. Deep Learning for Automatically Detecting Sidewalk Accessibility Problems Using Streetscape Imagery. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 196–209. doi:10.1145/3308561.3353798

[122] Thomas Westin. 2004. Game accessibility case study: Terraformers–a real-time 3D graphic game. In *Proceedings of the 5th International Conference on Disability, Virtual Reality and Associated Technologies, ICDVRAT*, Vol. 1. 95–100. https://www.researchgate.net/publication/260853226_Proceedings_of_the_5th_International_Conference_on_Disability_Virtual_Reality_and_Associated_Technologies_ICDVRAT_2004/links/0deec53282e0b15936000000/Proceedings-of-the-5th-International-Conference-on-Disability-Virtual-Reality-and-Associated-Technologies-ICDVRAT-2004.pdf#page=121

[123] Candace Williams, Lilian de Greef, Ed Harris, Leah Findlater, Amy Pavel, and Cynthia Bennett. 2022. Toward supporting quality alt text in computing publications. In *Proceedings of the 19th International Web for All Conference* (Lyon, France) *(W4A '22)*. Association for Computing Machinery, New York, NY, USA, Article 20, 12 pages. doi:10.1145/3493612.3520449

[124] Michele A. Williams, Amy Hurst, and Shaun K. Kane. 2013. "Pray before you step out": describing personal and situational blind navigation behaviors. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 28, 8 pages. doi:10.1145/2513383.2513449

[125] Jeff Wilson, Bruce N. Walker, Jeffrey Lindsay, Craig Cambias, and Frank Dellaert. 2007. SWAN: System for Wearable Audio Navigation. In *2007 11th IEEE International Symposium on Wearable Computers*. 91–98. doi:10.1109/ISWC.2007.4373786

[126] World Wide Web Consortium (W3C). 2019. *Images Tutorial - Web Accessibility Tutorials*. Web Accessibility Initiative (WAI). https://www.w3.org/WAI/tutorials/images/ Part of Web Accessibility Tutorials, providing guidance on how to create websites that meet WCAG.

[127] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1180–1192. doi:10.1145/2998181.2998364

[128] Paul Bach y Rita, Mitchell E. Tyler, and Kurt A. Kaczmarek. 2003. Seeing with the Brain. *International Journal of Human–Computer Interaction* 15, 2 (2003), 285–295. doi:10.1207/S15327590IJHC1502_6

[129] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything V2. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 21875–21911. https://proceedings.neurips.cc/paper_files/paper/2024/file/26cfdcd8fe6fd75cc53e92963a656c58-Paper-Conference.pdf

[130] Koji Yatani, Nikola Banovic, and Khai Truong. 2012. SpaceSense: representing geographical information to visually impaired people using spatial tactile feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 415–424. doi:10.1145/2207676.2207734

[131] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know? arXiv:2305.18153 [cs.CL] https://arxiv.org/abs/2305.18153

[132] Limin Zeng, Mei Miao, and Gerhard Weber. 2014. Interactive Audio-haptic Map Explorer on a Tactile Display. *Interacting with Computers* 27, 4 (02 2014), 413–429. doi:10.1093/iwc/iwu006 arXiv:https://academic.oup.com/iwc/article-pdf/27/4/413/9644221/iwu006.pdf

[133] Haixia Zhao, Catherine Plaisant, Ben Shneiderman, and Jonathan Lazar. 2008. Data Sonification for Users with Visual Impairment: A Case Study with Georeferenced Data. *ACM Trans. Comput.-Hum. Interact.* 15, 1, Article 4 (May 2008), 28 pages. doi:10.1145/1352782.1352786

[134] Kaixing Zhao, Sandra Bardot, Marcos Serrano, Mathieu Simonnet, Bernard Oriola, and Christophe Jouffrais. 2021. Tactile Fixations: A Behavioral Marker on How People with Visual Impairments Explore Raised-line Graphics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. doi:10.1145/3411764.3445578

[135] Kaixing Zhao, Rui Lai, Bin Guo, Le Liu, Liang He, and Yuhang Zhao. 2024. AI-Vision: A Three-Layer Accessible Image Exploration System for People with Visual Impairments in China. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 145 (Sept. 2024), 27 pages. doi:10.1145/3678537

[136] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173690

[137] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14.

doi:10.1145/3290605.3300341

[138] Yuhang Zhao, Elizabeth Kupferstein, Brenda Veronica Castro, Steven Feiner, and Shiri Azenkot. 2019. Designing AR Visualizations to Facilitate Stair Navigation for People with Low Vision. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 387–402. doi:10.1145/3332165.3347906

[139] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing Through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[140] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. 2018. Visual to Sound: Generating Natural Sound for Videos in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009