NaviNote: Enabling In-situ Spatial Annotation Authoring to Support Exploration and Navigation for Blind and Low Vision People

Ruijia Chen*
Department of Computer Sciences
University of Wisconsin-Madison
Madison, Wisconsin, USA
ruijia.chen@wisc.edu

Yuheng Wu*
Department of Computer Sciences
University of Wisconsin-Madison
Madison, Wisconsin, USA
yuheng.wu@wisc.edu

Charlie Houseago
Research and Development
Niantic Spatial, Inc.
London, United Kingdom
charlie@nianticspatial.com

Filipe Gaspar
Niantic Spatial, Inc.
London, United Kingdom
filipe@nianticspatial.com

Filippo Aleotti
Niantic Spatial, Inc.
London, United Kingdom
filippo@nianticspatial.com

Dorian Gálvez-López
Niantic Spatial, Inc.
Zaragoza, Spain
dorian@nianticspatial.com

Oliver Johnston
Niantic Spatial, Inc.
London, United Kingdom
ojohnston@nianticspatial.com

Diego Mazala
Niantic Spatial, Inc.
London, United Kingdom
diego@nianticspatial.com

Guillermo Garcia-Hernando
Niantic Spatial, Inc.
London, United Kingdom
guillermo@nianticspatial.com

Maryam Bandukda
UCL Interaction Centre
University College London
London, United Kingdom
m.bandukda@ucl.ac.uk

Gabriel Brostow
Department of Computer Science
University College London
London, United Kingdom
brostow@cs.ucl.ac.uk

Jessica Van Brummelen
Niantic Spatial, Inc.
London, United Kingdom
jess@nianticspatial.com

Figure 1: NaviNote enables blind and low vision (BLV) users to explore their surroundings via a five-stage pipeline after localizing their precise positions using Visual Positioning System (VPS) in a pre-scanned area on a smartphone: (1) User asks about the area using natural language; (2) User navigates to a specific location within the area by following the turn-by-turn instructions NaviNote provides; (3) User listens to nearby spatial annotations created by other BLV users; (4) User asks follow-up questions on spatial annotations and from other information sources; (5) User creates their own spatial annotations. Arrows in the figure indicate the internal flow of the pipeline. During the interaction, the user wears a vest that holds the smartphone with camera facing forward.

Abstract

GPS and smartphones enable users to place location-based annotations, capturing rich environmental context. Previous research demonstrates that blind and low vision (BLV) people can use annotations to explore unfamiliar areas. However, current commercial systems allowing BLV users to create annotations have never been evaluated, and current GPS-based systems can deviate several meters. Motivated by high-accuracy visual positioning technology, we first conducted a formative study with 24 BLV participants to

*Authors contributed equally to this research. Order was determined alphabetically.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790589

envision a more accurate and inclusive annotation system. Surprisingly, many participants viewed the high-accuracy technology not just as an annotation system but also as a tool for precise last-few-meters navigation. Guided by participant feedback, we developed NaviNote, which combines vision-based high-precision localization with an agentic architecture to enable voice-based annotation authoring and navigation. Evaluating NaviNote with 18 BLV participants showed that it significantly improved navigation performance and supported users in understanding and annotating their surroundings. Based on these findings, we discuss design considerations for future accessible annotation authoring systems.

CCS Concepts

• **Human-centered computing** → **Accessibility technologies**; **Accessibility systems and tools**; **Accessibility design and evaluation methods**; **Mixed / augmented reality**.

Keywords

Accessibility, Blind and Low Vision, Last-Few-Meters Navigation, Spatial Annotation, Augmented Reality, Visual Positioning System, Large Language Models

ACM Reference Format:

Ruijia Chen, Yuheng Wu, Charlie Houseago, Filipe Gaspar, Filippo Aleotti, Dorian Gálvez-López, Oliver Johnston, Diego Mazala, Guillermo Garcia-Hernando, Maryam Bandukda, Gabriel Brostow, and Jessica Van Brummelen. 2026. NaviNote: Enabling In-situ Spatial Annotation Authoring to Support Exploration and Navigation for Blind and Low Vision People. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3772318.3790589

1 Introduction

Exploring and understanding one's surroundings is essential for human well-being [31, 126] yet remains a persistent challenge for blind and low vision (BLV) people. Large language models (LLM) and computer vision (CV) capabilities have supported BLV users with in-situ exploration by detecting nearby objects [56, 72, 87, 93], recognizing text [53, 117], and providing scene descriptions [44, 45, 45, 65, 79, 89, 113, 123], yet these camera-based AI tools impose limitations for BLV users around aiming the camera correctly [16, 23], verifying information, and addressing LLM hallucinations (i.e., models generate plausible-sounding but unfaithful or nonsensical information [58]) [3, 50, 149]. They also cannot provide information about objects outside the camera's frame.

Location-based annotations complement existing tools by providing verified contextual information outside the camera view (e.g., an entrance behind the user). Prior work, Footnotes [36], has shown their value in providing multifaceted information to BLV users that LLM- and CV-based tools alone cannot provide. However, BLV users play a passive role in FootNotes as well as other assistive systems for exploration [25, 33, 60], consuming information along predetermined routes rather than contributing their own situated knowledge. Existing systems are also limited by the precision of localization technologies, often relying on the 10-meter accuracy of GPS [86] or similar, which limits the utility of annotations as they cannot be precisely placed or accessed.

New emerging technologies could bridge this gap, allowing the creation of accessible, empowering interfaces for BLV users to act as active in-situ explorers and annotation contributors. First, multi-modal large language models (MLLM) can enable voice-first interfaces for users to interact accessibly with their surroundings [60]. Second, the latest Visual Positioning Systems (VPS) are far more precise than GPS at tracking where the user's phone is in 3D space [52, 104], without needing to aim cameras at specific targets. We propose that together, these technologies offer opportunities for high-accuracy, conversational, in-situ spatial annotations. We therefore seek to answer the following research questions in the context of such a system:

- RQ1: What types of annotations are most valuable to BLV users, and in what scenarios are they most applicable?
- RQ2: How can we design an accessible interface to enable BLV users to author in-situ spatial annotations?
- RQ3: How can BLV users effectively access and utilize spatial annotations?

To answer these questions, we first conducted a formative study involving 24 BLV participants, identifying six key annotation categories: *Safety*, *Accessibility*, *Amenity*, *Layout*, *Attraction*, and *Experience.* Beyond annotations, we learned our proposed technology could address a critical prerequisite for exploration: precise "last-few-meters" navigation, providing turn-by-turn instructions during the final meters of travel where traditional GPS often fails [127]. Motivated by these findings, we designed *NaviNote*, a voice-based system combining VPS technology with an AI-powered conversational interface. Utilizing sub-meter-level localization (i.e., position error of less than one meter), NaviNote provides turn-by-turn instructions and an audio compass to facilitate safe navigation for BLV users, and supports in-situ accessing and authoring of annotations at precise locations.

To evaluate NaviNote, we conducted a second study with 18 BLV participants in a local public square, now focusing on navigation and free exploration. Sixteen participants compared navigation using NaviNote versus a baseline, and all 18 participants completed a free-form exploration to listen to pre-set annotations and create new ones. We found NaviNote significantly improved navigation performance, with 14 of 16 participants successfully navigating using NaviNote compared to only six with the baseline, and also enabled all participants to successfully create annotations independently. Moreover, NaviNote significantly outperformed the baseline across qualitative measures, including perceived effectiveness, usability, mental demand, frustration, and perceived performance. Based on participants' in-situ authoring, we refined the taxonomy of desired annotations and introduced another category, *Request.*

In summary, our contributions include: (1) NaviNote, a voice-based system that enables BLV users to create in-situ spatial annotations and navigate effectively in last-few-meters scenarios, (2) an evaluation of NaviNote with 18 BLV participants, demonstrating its benefits for navigation and exploration, and (3) a taxonomy of BLV users' desired spatial annotations, developed through two studies (n_1=24, n_2=18).

2 Related Work

We situate our system's initial design within prior work on supporting BLV users in understanding their surroundings, navigating the last-few-meters, and using in-situ spatial annotations, as well as on crowdsourcing for accessibility.

2.1 User-Initiated Tools for Environmental Understanding

Assistive technologies often help BLV users understand their surroundings through user-initiated requests. Many of these are camera-based, supporting magnification [134, 144], object recognition [56, 87, 93], and text recognition [53, 117] in real time. However, properly aiming a camera is often difficult for BLV users [51, 57, 140], and the information returned is typically limited to describing each individual object or reading text without providing broader contextual cues needed for navigation and orientation [50].

Several tools have introduced conversational AI to improve BLV users' access to environmental information. Seeing AI [89] and Be My Eyes [143], two widely used smartphone apps [34], let users capture photos and receive detailed audio descriptions. Research prototypes push this further: ChitChatGuide [60] supports question answering about nearby points of interest (POIs) using LLMs; WorldScribe [16] generates real-time visual descriptions tailored to users' contexts (e.g., concise descriptions for dynamic scenes, detailed descriptions for stable settings); and StreetViewAI [33] enables virtual exploration of destinations via LLM-accessible streetscape maps. These systems move beyond recognition to enable richer, context-sensitive queries. Yet they still primarily position BLV users as passive consumers: while users can ask questions and receive descriptions, they cannot contribute new, in-situ knowledge to the location.

To our knowledge, no tools currently enable BLV users to share situated knowledge and lived experiences about the physical world in-situ. Our research addresses this gap.

2.2 In-Situ Spatial Annotations and "Last-Few-Meters" Navigation

Spatial annotations bind information to locations, providing answers beyond camera-based recognition. Prior work has described this as creating "chatty environments," where the environment "speaks" to users to provide otherwise inaccessible information [22]. For sighted users, spatial annotations have long been studied. For example, GeoNotes [114] enabled users to attach notes to specific locations using PDAs, ActiveCampus [40] allowed students to share annotations to discover labs and nearby events, and StoryPlace.Me [9] explored how elderly users could leave location-based stories for future generations.

A few spatial annotation systems have targeted BLV users [23–25, 36, 84, 133], with FootNotes [36] being most relevant. In FootNotes, researchers attached annotations to existing OpenStreetMap objects, which would play when BLV participants' GPS signals were nearby. Yet, the system did not provide an interface designed specifically for BLV users to author annotations, and rather assumed screen readers and text-to-speech would suffice without user evaluation. Several GPS-based commercial systems, such as Sendero BrailleNote GPS [74], Trekker Breeze [32], BlindSquare [91], and Microsoft Soundscape [88], allowed BLV users to actively mark their environments. Users can search for and learn about important nearby POIs (e.g., crossroads, stores, popular cafes), mark their own POIs, and save routes for later retracing via voice interaction. However, these GPS-based systems, including FootNotes and commercial tools, face the key constraint of the "last-few-meters" problem [30, 36, 75, 127], where GPS fail to provide accurate turn-by-turn guidance in the final few meters—sometimes as big as tens of meters [92]—due to signal inaccuracy or incomplete map data [127]. As a result, BLV users fail to accurately locate and reach the intended POIs. Even when BLV users know they are near or directly facing a POI, they often lack adequate information (e.g., haptic textures, smells, or characteristic sounds for landmark recognition [76, 127]) to confidently use the location with GPS.

To bridge this gap, some systems have introduced infrastructure for higher accuracy, such as Bluetooth beacons (e.g., NavCog3 [130]), NFC-tagged landmarks [35], or ultrawide-bandwidth beacons [77], but these require costly hardware installations and pre-built floor maps. Others have explored navigation without pre-built maps [64, 65, 127, 148]. For example, WanderGuide [65] allowed users to explore the environment with a suitcase-shaped robot, providing users with walkable directions and customizable levels of audio scene descriptions. Clew [148] enabled users to record and later retrace routes using smartphone sensors. Closing the Gap [127] demonstrated the potential of camera-based landmark recognition to assist users near a destination, but relied on a Wizard-of-Oz setup [26] rather than a deployable pipeline. None of these systems provided BLV users with a low-cost, low-barrier-to-entry, independent solution to the last-few-meters problem.

Visual Positioning Systems (VPS) provide a new opportunity for the last-few-meters navigation at high accuracy. In addition to GPS information, VPS combine input from multiple sensors (e.g., camera, motion sensors, depth sensors) to generate visual features in the surrounding environment. As users move their phone and scan nearby objects or buildings, VPS compare what the camera sees with a large database of previously mapped visual features to pinpoint the user's exact location and orientation with sub-meter accuracy [38, 68, 116, 122]. Motivated by the gaps in accessible annotation authoring and precise last-few-meters navigation, NaviNote tackles these dual challenges with natural language interaction and VPS–based localization to empower BLV users to navigate to, access, and author their own spatial annotations independently using a smartphone.

2.3 Crowdsourcing for BLV Users

Crowdsourcing annotations has been widely used to make locations more accessible for BLV users. Previous research explored online crowdsourcing to annotate bus stops [46, 118], accessible sidewalks [47, 70, 128], storefronts [73], obstacles [124], semantic information in the real world environment [37, 43], and alt text for online images [129]. Other work demonstrated how crowdsourced descriptions of artwork can enhance museum experiences for BLV users [115].

In addition to large-scale labeling, several systems connect BLV users with sighted volunteers for on-demand assistance. For example, Be My Eyes [143] allows BLV users to share their camera view with sighted helpers for real-time support, while VizWiz::LocateIt

[10] enables BLV users to post photos and questions online, allowing remote workers to outline objects and help them locate targets.

Despite these advances, most crowdsourcing tools cast BLV users as requesters rather than independent contributors. Yet BLV users hold valuable perspectives, which, when shared, can better address their communities' needs [50]. As AI becomes increasingly capable of automatic annotation, it is critical to identify what BLV users value most, rather than relying on what AI "decides" is important. Our two studies investigate these values and needs, and NaviNote empowers BLV users to act as active information contributors.

3 Formative Study (Study 1)

To inform the design of NaviNote, we conducted a formative study with 24 BLV participants, focusing on whether they would value highly-localized audio annotations, what they might use them for, and where they might place them (RQ1). The study identified BLV participants' desired types of annotations and revealed how precise navigation would be a prerequisite for annotation authoring.

3.1 Methods

3.1.1 Participants (n_1=24). We recruited 24 BLV participants (F1-F24) from the local BLV communities. Table 1 in Appendix A lists participant demographics and visual conditions. Participants received a compensation of 75 GBP.

3.1.2 Procedure. We conducted a user-centric design workshop, which consisted of two identical sessions over two days with 12 participants each day. In each session, participants were randomly assigned to one of three four-person groups. Each session included four stages: scenario-based brainstorming, tactile-map prototyping, Wizard-of-Oz (WoZ) authoring [26], and reflection. Both sessions lasted for approximately three hours.

Scenario-based Brainstorming. Inspired by the co-design activity in [15], we introduced participants to seven different scenarios (see Supplementary Section **??**), where real-world spatial annotations with centimeter-level accuracy could be useful. Participants then voted and held small-group discussions on the top two scenarios, which were as follows. **Day 1**: (1) What audio annotations could Orientation & Mobility (O&M) teachers add to a space to act as a training ground for you? (2) You're at home organizing your room. What annotations would you place to help with organization? **Day 2**: (1) You're traveling to a friend's apartment complex to house-sit for them. What annotations would you want them (or the complex) to place ahead of time? (2) You're making a location-based audio game. What annotations would you place?

Tactile-Map Prototyping. We provided each group with a 3D-printed tactile map of a well-known local area, e.g., a historical site, a large intersection, or a park with a garden. Participants audio-recorded annotations, and placed markers on the map indicating the annotation locations (see Figure 2). They prototyped for approximately 45 minutes and described their prototypes to the larger group.

WoZ Authoring. We conducted a WoZ study on UCL East campus, where one researcher acted as an AI system that participants could ask questions of—e.g., about the surroundings or to place an audio note—and another researcher acted as the location-based audio playback system using an audio recorder. Each group engaged with the WoZ separately for 15 minutes, authoring their own annotations indoors and outdoors on campus.

Reflection. Participants engaged in semi-structured discussion for 15-30 minutes in a large group about the pain points of the proposed system, how they envisioned the system's form factor, and the most and least important system features. (See Supplementary Section **??** for full list of discussion questions.)

3.2 Analysis

We audio-recorded all study sessions, transcribed them using Whisper [120], and manually fixed transcription errors. Transcripts were analyzed using thematic analysis [14, 20]: Two researchers open-coded the transcripts from the first day of the workshop (approximately 50% of the data) independently, and developed a codebook by discussing their codes to resolve disagreements. The two researchers then split the transcripts from the second day, coded independently based on the initial codebook, and added new codes when necessary. During this process, the two researchers periodically discussed new codes to ensure consistency. New codes were added to the codebook upon agreement. The research team then developed themes and sub-themes from the codes by clustering relevant codes using axial coding and affinity diagrams [138]. Finally, the researchers cross-referenced the original data, the codebook, and the themes to make final adjustments, ensuring all codes were correctly categorized. For this formative analysis, our emphasis is on the range of perspectives rather than numerical prevalence [83]. We present them as motivation and grounding for our system design.

3.3 Findings

Here, we summarize participants' motivation for spatial annotations and present six categories of annotations they desired. We also present their preferred system features to derive design guidelines for NaviNote.

3.3.1 Annotation Motivation. Participants identified two primary purposes for spatial annotations. First, they hoped to use annotations to share and access information within their social circles (e.g., facilities in their home when their friends were visiting) and the broader BLV community (e.g., safety warnings for other BLV people). Second, participants envisioned documenting in-situ experiences, either for personal use (e.g., revisiting and recollecting memories) or for collective benefit (e.g., sharing comments and feedback about the accessibility of locations).

Beyond location-based annotations, participants also described potential needs for dynamic object-based annotations, primarily for indoor scenarios, such as locating and organizing household items (e.g., matching clothing, identifying spices or cleaning products), leaving operational instructions or reminders of food expiration dates for safety, and attaching personal memories to specific objects like photos. While equally important, we focus on location-based annotations in this work and leave dynamic object annotations for future exploration.

3.3.2 Annotation Taxonomy. We identified six categories of desired spatial annotations for BLV users: safety, accessibility, amenity,

Figure 2: Left: One of the formative study small groups prototyping where they would place annotations on a 3D-printed map. Right: An example of a final prototype, with various audio buttons—which contain annotation recordings—arranged around a tactile map. The orange modeling clay indicates annotation locations.

layout, attraction, and experience. We noticed these categories often overlapped with those from FootNotes [36], but also extended them—e.g., our "Attraction" category broadens FootNotes' narrower focus on "Historical"—reflecting the wider range of annotations found in participant quotes.

Safety. Participants emphasized safety-related annotations, including depth-related (e.g., stairs, curbs, uneven pavement, cracks) and width-related features (e.g., narrow entrances). They were attentive to height, width, orientation, and number of steps due to challenges in depth perception. Furthermore, they wanted annotations near water (e.g., rivers, fountains, low-fenced areas) to alert them of risks of falling, and annotations about protruding or sharp obstacles such as hanging branches and thorns. Participants also highlighted crossroads as important to be annotated, including the position of zebra crossings, curbs again, and "where the traffic lights (buttons) are." (F2)

Accessibility. Participants noted a need for annotations about accessible facilities (e.g., elevators, escalators, ramps). For example, they proposed annotations distinguishing accessible toilets (i.e., toilets specially designed to accommodate people with disabilities, such as those equipped with braille signage and offering larger space) from generic ones to help find them faster. Participants also wanted annotations that identified "haptic paths," as F15 described, "something with a touch and tail." They wanted to annotate features that could be held or followed with a white cane, such as handrails, walls, or curbs. Participants also wanted to annotate help points (e.g., reception desks and concierges). As F2 noted: "You may need to know where the reception desk is, so you can speak to someone."

Amenity. Participants wanted annotations about amenities, including public transportation (e.g., train stations, bus stops), common functional areas (e.g., shops, parks, cafes, prayer rooms), and facilities (e.g., toilets, bins, benches). They were also interested in information related to such amenities, such as opening and closing times, public access, and peak hours of use.

Layout. Consistent with prior work [17], besides annotations on individual locations, participants also desired an overview of their surroundings, such as the overall shape and size of their current area, locations of entrances and exits, and the start and end of passageways. They also wanted information about the boundaries of different areas, such as the boundary between pedestrian zones and bike lanes.

Attraction. Participants wanted to annotate tourist attractions, such as providing historical details about a location, similar to QR code-based tours in museums. They also wanted annotations that shared information about upcoming events, such as exhibitions nearby, and envisioned annotating salient landmarks (e.g., a big tree) to support mental map construction and to share as meeting points.

Experience. Finally, participants wanted to create annotations documenting personal experiences and comments about particular places. As F3 noted: "It would be nice to have a note, like, oh, 'I came here with my friend, and we went to this building.' " Participants also suggested that such experience annotations could serve as a community feedback channel, similar to "an audio version of Google Maps reviews" (F4) [39], where they can provide suggestions to service providers for improving accessibility. For example, F23 suggested that cafes should offer assistance to people with additional needs when making purchases or payments upon entering the store.

3.3.3 Design Guidelines. Participants shared their desired features for an annotation-authoring system. We summarize their needs into six design guidelines (DG1-6) for NaviNote. Several of these guidelines also align with findings from prior work [33].

DG1: Support Navigation with Constant Confirmation. Participants across groups in the Reflection activity agreed that a wayfinding feature would be essential: Why have annotations if you can't independently explore the world?

DG2: Support Different Levels of Conversational Queries. Participants highlighted how the system should answer questions about objects in front of them, and provide higher-level information about their surroundings. This could include recommendations about nearby places or concise summaries of the area. As F23 detailed, "I was thinking, 'What if they have something all in one, like a kind of a introduction of anywhere, and it will tell you what you expect. What's in front of you? When you turn there?' It's like a two minute [introduction]."

DG3: Play Certain Annotations Automatically, Especially Safety. Participants suggested some annotations should be played automatically, noting that even seemingly irrelevant information might turn out to be useful. As F24 said, "We should never fathom that it was there, because if you don't switch it on, you don't know, right? If you just point at something you want and you don't realize what's around you." All participants noted safety annotations as most important and that they should be played automatically, possibly alongside vibrations or beeping alerts.

DG4: Allow for Information Filtering. While agreeing that certain annotations should be played actively, participants also expressed wanting annotation filters, which could be customized based on visual abilities and preferences. As F2 explained: "If I'm in an area that I'm fairly familiar with, I may need less [information] layers to make sure I'm walking straight. Whereas if I'm in a completely new area that is unfamiliar to me, I may want a bit more information." F7 added, "During the day, I actually don't need that much information. And then when it suddenly goes dark... I need more information. Then it's like, 'I nearly tripped.' "

DG5: Enable BLV Accessibility and Customizability. Participants stressed that the system must be accessible and adaptable to diverse needs, suggesting the use of voice input and tactile buttons. In the Tactile-Map Prototyping phase, large physical buttons (Figure 2, right) were used to simulate recording audio annotation, inspiring participants to recommend similar tactile buttons alongside voice input for more accessible control in the final design.

Given VPS requires camera information, participants suggested attaching it to the body to simplify aiming. Moreover, they highlighted the importance of customization, such as adjusting voice settings, and navigation directions (e.g., clock-face vs. left-and-right), which aligns with prior findings on the need for personalization [16, 33, 50].

DG6: Offer Multi-modal Feedback. Participants expected the system to offer both haptic and audio feedback. For example, vibrations could be used for obstacle warnings, while spatial audio could guide users toward destinations. Visual feedback was considered important for low-vision users, as they relied heavily on residual visuals [136]. Interestingly, blind participants also wanted visual feedback for sighted companions who might assist them. As F7 explained, "I'm surrounded with people with full vision. And sometimes when it's all audio, they're like, 'I can't help you'. So I feel like for the sighted people that are going to assist me, they do need a visual cue."

4 NaviNote: Voice-based Annotation and Navigation System for BLV Users

Based on the formative study findings, we developed an interaction framework to enable BLV users to access spatial annotations independently (RQ2). We then designed and implemented NaviNote, the first system that enables BLV users to conversationally create and access in-situ annotations. NaviNote is hands-free and voice-based, and supports querying the surroundings, navigating to objects or annotations of interest, and creating, editing and deleting annotations. We included customization features to allow users to tailor AI responses and UI to their needs.

4.1 Interaction Pipeline for Annotation Authoring

Our five-stage interaction pipeline for accessible exploration and annotation authoring follows:

Scanning. NaviNote relies on pre-scanned and pre-processed environment data from a Point of Interest (POI) (i.e., a geographic area of interest [135]) to accurately determine users' pose (i.e., position, orientation) and understand scene elements. Relying on pre-scanned data removes the burden for BLV users to properly scan the environment in real time [67] or aim the camera precisely [3, 16]. Various other research and industry applications utilize scans in the same way, including ImaginateAR [67], CoCreatAR [105], Google Geospatial API [38], and Niantic Wayfarer for Pokémon GO [99], and have shown that crowdsourcing scans can effectively gather wide, accurate coverage [97, 107]. We envision this pre-scanning process being performed by sighted users, volunteers [99], or even other BLV users [139]. We discuss the implementation details of pre-scanning and pre-processing in Section 4.2.1.

Localization. Upon launching NaviNote, users perform a brief localization process (5-10 seconds) by slowly moving their devices horizontally and vertically to observe the surrounding environment. The system detects visual features and to determine the user's pose with sub-meter-level accuracy. Once localized, NaviNote fetches the associated spatial data for navigation and annotation.

Query. Users can ask NaviNote questions about their surroundings or other topics of interest (e.g., history of a nearby statue). NaviNote provides responses by synthesizing information from multiple sources, including the pre-scanned environment, existing spatial annotations, users' camera feed, public map sources, and the internet.

Navigation. When asked, NaviNote provides turn-by-turn instructions to guide users to target objects or annotations, and provides hazard warnings en route drawn from safety-related annotations.

Annotation Authoring. Users can create annotations at specific locations by conversing with the system, such as by saying "Place an annotation on this bench saying it has armrests." They can also edit or remove existing annotations via a similar conversational procedure.

4.2 System Design and Implementation

Based on the interaction pipeline and formative study, we designed NaviNote with four major components: Scanning and Localization, Query Answering, Navigation, and Annotation Accessing and Authoring. NaviNote contains a phone-based frontend for user interaction and a Python server backend for request processing. We built the frontend with Unity 6000.0.34f1 and Niantic SDK for Unity [103], and deployed the prototype on an iPhone 14 Pro with iOS 18.6.1. Regarding the form factor, we used a commercial running vest with adhesive straps to secure the phone on the users' chest for hands-free operation (e.g., while carrying a white cane). For audio output, we used L5 Bone Conduction Earphones, as bone conduction earphones enable users to hear system instructions while still perceiving environmental sounds simultaneously [7]. Wired earphones were also provided as an alternative.

4.2.1 Scanning and Localization.

For NaviNote to dynamically track users' pose and understand their surrounding environment, we first scan the whole POI to generate its 3D map. We adopt Niantic's VPS framework [104] both when making and using our bespoke NaviNote map, as this framework supports crowdsourced scans from commercial devices with depth sensors (e.g., recent iPhone or Android phone equipped with LiDAR) and provides high coverage of popular POIs [98]. Users conduct the scanning by pointing their devices at the surroundings, while the framework captures synchronized multi-modal sensor streams including visual imagery, depth information, and device poses. In our experiment setup, scanning the study area of approximately 45 m × 45 m takes 22 minutes. These streams are aggregated and uploaded into Niantic's VPS server where the data is transformed into spatial features for localization in future sessions.

Besides spatial features, we also construct a *scene graph* of each POI from its 3D map. A scene graph describes the spatial properties (i.e., position, rotation, and dimensions, represented as bounding boxes in the 3D point cloud) of each object within the POI. Similar to the scanning process, the scene graph can be generated through crowdsourcing by sighted users and volunteers. Prior work also demonstrated the practicality of automatically generating scene graphs from RGB-D inputs [41, 78, 142]. In our user evaluation study, we manually annotated the scene graph of the experimental scene to ensure accuracy.

Ultimately, when a user is running NaviNote, the system loops, estimating the user's current pose relative to the scene graph and fetching spatial annotations from the backend, and providing these as context to NaviNote's conversational agent system.

4.2.2 Query Answering.

One of NaviNote's most notable features is its capability to answer user queries with natural language, as illustrated in the top panel of Figure 3. This was motivated by DG2 and implemented in the "Orchestrator" pattern [110] with MLLM agents. The backend contains a single entry point (i.e., the orchestrator) for user queries and has access to modules specialized for particular purposes (i.e., agents). The orchestrator analyzes user queries, delegates them to agents with suitable tools to gather relevant information, and synthesizes a final response.

We implemented six agents, each designed to handle different types of user queries, supporting real-time understanding of the surroundings through the camera and external databases (e.g., our pre-defined annotations and public map data). **Annotation Agent**, which retrieves, creates, edits, and removes spatial annotations. It answers questions related to existing annotations (e.g., "What do people think of this park?"); **Scene Understanding Agent**, which interprets scene graphs of POIs and computes object positions to answer spatial queries (e.g., "What's in front of me?"), and triggers navigation on demand; **Image Search Agent**, which accesses users' camera frames and answers visual questions (e.g., "What color is this flower?", "Please read the sign."); **Internet Search Agent**, which retrieves information from the internet; **Public Map Agent**, which uses public map data (e.g., Open Street Map [111]) to gather geographic information of nearby areas at a larger scale with coarser granularity (e.g., nearest train station) to supplement the scene graphs; and **Customization Agent**, which interprets user customization requests and sends them to the frontend for execution. All agents have access to users' pose provided by VPS from the frontend. We used OpenAI GPT-4o [108] for the orchestrator and GPT-4o mini [109] for the agents.

The backend is hosted on a cloud machine and communicates with the frontend via WebSockets [29]. Users query by speaking to the frontend, which transcribes queries locally using Whisper [120] and forwards the transcribed text to the orchestrator. The orchestrator invokes relevant agents and triggers actions on the frontend as needed, and finally aggregates agents' outputs into a response, which is converted to audio on the frontend via iOS text-to-speech API. The orchestrator is instructed to give responses within two sentences or approximately 45 words by default, and users can customize its verbosity. Users can also provide their visual conditions or desired information (e.g., "include color information") to help the orchestrator produce more helpful responses. We provide the system prompt in Supplementary Section **??**.

Following participants' preference for physical buttons (DG5), we reprogrammed the physical volume button to serve as the interaction trigger. To issue a query, the user presses the volume button (either on the smartphone or on connected earphones), speaks their question, and waits for the system to process the input.

4.2.3 Navigation.

NaviNote supports last-few-meters navigation to facilitate scene exploration, as illustrated in the middle panel of Figure 3. As the scene graph only contains information about static objects within the POI and does not reflect the environment's dynamics (e.g., people nearby), NaviNote uses the Semantic Mesh provided by Niantic SDK [102] to construct a real-time mesh of walkable ground. We further refine the walkable mesh by excluding parts that collide with objects in the scene graph. Then, NaviNote computes the shortest path from the user to the destination using the A-star algorithm [48]. The path is then smoothed to retain only key turning points. If the path aligns with the edges of objects in the scene graph, it forms a "haptic path" (Section 3.3.2) and generates navigation instructions accordingly (e.g., "Correct direction. Follow the edge of the flower bed on your right").

NaviNote delivers navigation instructions in two ways: turn-by-turn instructions and audio compass. Turn-by-turn instructions are announced in a concise format (e.g., "2 o'clock, 5 meters"). To provide timely feedback of direction (DG1), they are triggered at turning points and every 10 seconds as confirmation (e.g., "Correct direction, 2 meters"), and when the user deviates from the planned route. The distance to the destination is also updated regularly (e.g., "Destination lies at 12 o'clock"). All audio instructions are played in spatial audio to provide directional cues [61, 62]. Additionally, we developed an "audio compass" to provide users with immediate and constant feedback on their direction (DG1). The audio compass is a beeping signal that continuously plays at high volume when the user is facing the correct direction and at low volume when not (see Figure 4). Following DG6, we also implemented visual feedback, including a green-line path and a star marker at the destination (see Figure 5).

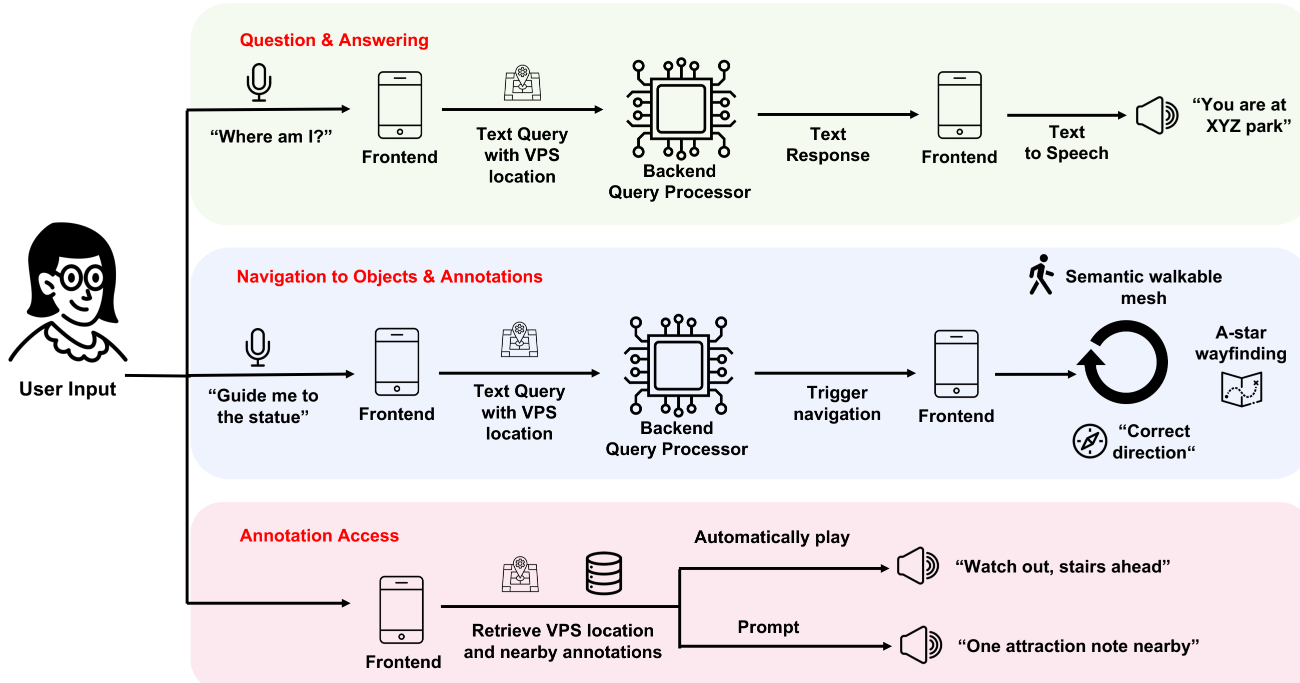

Figure 3: The interaction flow of NaviNote. Top: Question answering. The user asks a question to the phone, which converts speech to text and sends the query with VPS location to the backend. The backend returns a text response, which the frontend speaks aloud via text-to-speech. Middle: Navigation to objects and annotations of interest. The user requests guidance to an object or annotation. The query and location are sent to the backend, which triggers navigation on the frontend. As the user moves, the frontend generates turn-by-turn instructions with an audio compass with A-star wayfinding algorithm over the semantic walkable mesh. Bottom: Annotation access. The frontend continuously retrieves VPS location and nearby annotations, playing them either automatically or by prompt.

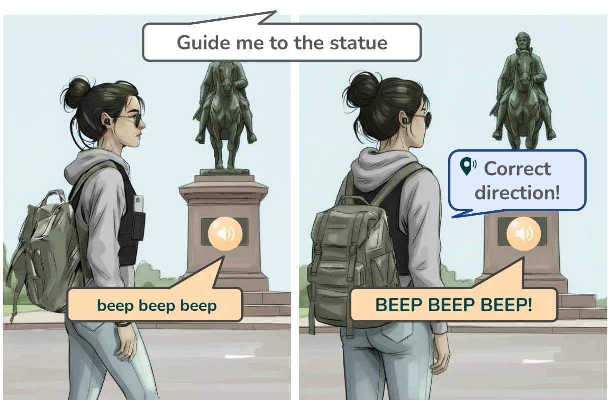

Figure 4: During navigation, there is an audio signal ("beep") in the direction users should be moving. When the user's orientation is correct, the audio signal continuously plays at high volume, whereas at a lower volume with an incorrect orientation. When the user's orientation is correct, the system also periodically provides verbal confirmation, saying, "Correct direction!"

NaviNote allows users to customize navigation instructions. They can choose among different directional formats, including clock-face (e.g., "Turn to 11 o'clock"), egocentric (e.g., "Turn forward-right"), egocentric with degrees (e.g., "Turn left by 30 degrees"), and cardinal (e.g., "Facing south. Head East"). Users can also select their preferred unit of distance: meters, feet, or steps (with one step defined as 0.76 meters).

4.2.4 Annotation Accessing and Authoring. NaviNote supports accessing existing spatial annotations (see the bottom panel of Figure 3) and authoring new ones. Spatial annotations are triggered when the user approaches within a predefined distance (1.5 meters for safety and accessibility annotations, and 1.0 meter for other less critical categories). Following DG3 and DG4, NaviNote provides two mechanisms for accessing annotations to provide layered access to avoid information overload. Different access mechanisms are applied based on annotation category: by default, *Safety* and *accessibility* annotations are **automatically played** to ensure timely delivery of critical information. For safety annotations, the frontend also vibrates for 0.5 seconds (DG6). The other four categories are accessed **only by asking**: when nearby, the system delivers a prompt with a short sound effect ("bing") followed by a notification aggregating these annotations (e.g., "two attraction notes nearby"). Users can ask follow-up questions to access them. When multiple annotations are nearby, annotations set to play automatically are prioritized over those set to be prompted. All messages are played sequentially, and any message that cannot be played within 15 seconds is skipped to avoid information overload. Users can also adjust which categories of annotations are automatically played, prompted, or ignored by talking with the system (see Section 5.2 for participants' choice).

For authoring, users can create annotations in natural language by simply describing the new annotation (e.g., "Place an annotation here saying, 'Exit ahead'.") Annotations can be placed on objects in the area or at the user's current location. The backend automatically determines the annotation category from the content and acknowledges successful creation. Users can also modify or remove annotations they previously authored through follow-up commands (e.g., "I want to delete the annotation about the exit"). We set all annotations to be publicly accessible for the convenience of prototyping, and leave the management of public and private annotations as future work (see more discussion in Section 7.1.3).

5 Evaluation (Study 2)

To evaluate the effectiveness of NaviNote in assisting BLV people in exploring outdoor environments (RQ3), we conducted a study with four pilot participants and 18 study participants in a public local square. We aim to answer the following questions: (1) What did BLV users ask NaviNote during outdoor exploration?, (2) How effective is NaviNote in supporting BLV users in last-few-meters navigation?, (3) What do BLV users think about the exploration features, including the process of annotation access and authoring, and the existing annotation types?, and (4) How do BLV users envision using NaviNote in other real-world scenarios?

5.1 Participants (n_2=18)

We recruited 18 BLV participants (11 females, 7 males) from local low-vision communities and via snowball sampling [112]. Their ages ranged from 20 to 80 ($Mean$ = 48.9, SD = 16.5). Six participants reported being totally blind and eight reported being legally blind. Table 2 in the Appendix B details participants' demographics and visual conditions. Eight participants (P1-P4, P6, P10, P12, P16) also took part in the formative study. Two participants did not complete the navigation task (Section 5.3): P17 withdrew by choice and P18 was interrupted externally. Both participants only experienced NaviNote during tutorial and the free exploration task. All 18 participants received compensation of 75 GBP.

We also recruited four BLV participants (see Table 2 in the Appendix) for pilot studies before the formal evaluation. After these pilot studies, we refined the experimental scripts to better explain system features and gathered valuable feedback from participants that informed our baseline condition choice (Section 5.2). We also fixed system engineering bugs, and adjusted the annotation trigger distances from 2 meters to 1.5 meters for safety and accessibility annotations, and to 1.0 meter for other categories (Section 4.2.4), to help users better locate annotations while still receiving safety alerts in advance.

5.2 Apparatus

We conducted the study in a local public square (Golden Square in London) selected for its size (~40 × 40 m), which represented a realistic navigation scenario, and for its safety, with fencing along the perimeter. The ground surface within the experimental area was uniformly flat and did not contain any designated tactually salient pathways, though some existing structures (e.g., curbs, flower beds) were naturally perceivable by touch. See Figure 6(A) for the square layout. We planned two routes for participants to navigate, each approximately 30 m long with an equal number of turns (i.e., two turns) [55, 84] and a similar number of landmarks (e.g., table tennis tables, flower beds) [90, 147], as shown in Figure 6(B). Since the square contains three large elevated flower beds with curbs in the middle, both designed routes required participants to walk around one of the flower beds. The research team created 39 annotations within the square, with 11 safety, 7 accessibility, 4 amenity, 6 layout and 4 experience annotations, which were all inspired by quotes and findings from the formative study, as shown in Supplementary Section **??**.

Given the lack of existing tools to address the last-few-meters problem for BLV people, we used photo describer tools (e.g., SeeingAI [89]) as baselines for navigation. Initially, we tested the Apple Magnifier [5] in a pilot study with four BLV participants. Participants reported that it caused information overload, due to its continuous stream of information. Pilot4 commented: "This is something too much to keep listening and listening," and Pilot1 added: "Because too much [information caused] information [over]load, you might think it's not [even] information." Pilot4 suggested TapTapSee [137] for the baseline, which we adopted, as it is commonly used in the BLV community. TapTapSee delivers concise, on-demand image descriptions by double tapping the screen. We considered audio-based tools as baseline, such as VoiceVista [145], which offers audio beacons for navigation. However, the beacons rely on GPS localization, which is not precise enough for last-few-meters navigation [86].

5.3 Procedure

The study consisted of a single on-site session that lasted approximately 2.5 hours. Prior to the on-site session, participants completed a demographics questionnaire (see Appendix B). The on-site session contained four parts: a tutorial on NaviNote, a navigation task, a free-form exploration, and an exit interview.

5.3.1 Tutorial. We started with a tutorial on NaviNote and the navigation baseline, TapTapSee. For NaviNote, the localization process was completed in advance by the research team to ensure consistent localization quality across all participants, allowing the later evaluation to begin from the same starting point. We first instructed participants to press the volume button of the earphone to ask questions. Fifteen participants used bone conduction earphones and three (P1, P3, P4) opted for wired alternatives. We demonstrated how to create, edit, and delete annotations using natural language, and how to access annotations either automatically or on demand. Participants also learned about the navigation function, including turn-by-turn instructions and the audio compass, through a short practice route that did not overlap with the later experiment. We ensured that both the high and low audio compass volumes were audible and comfortable for all participants. Participants were welcome to customize voice timbre (e.g., gender, accent) and speed, and set which type of annotations they wanted to be played automatically, prompted, or ignored. All participants kept *safety* and *accessibility* annotations set to play automatically, with the remaining categories set to be prompted. For TapTapSee, we explained how

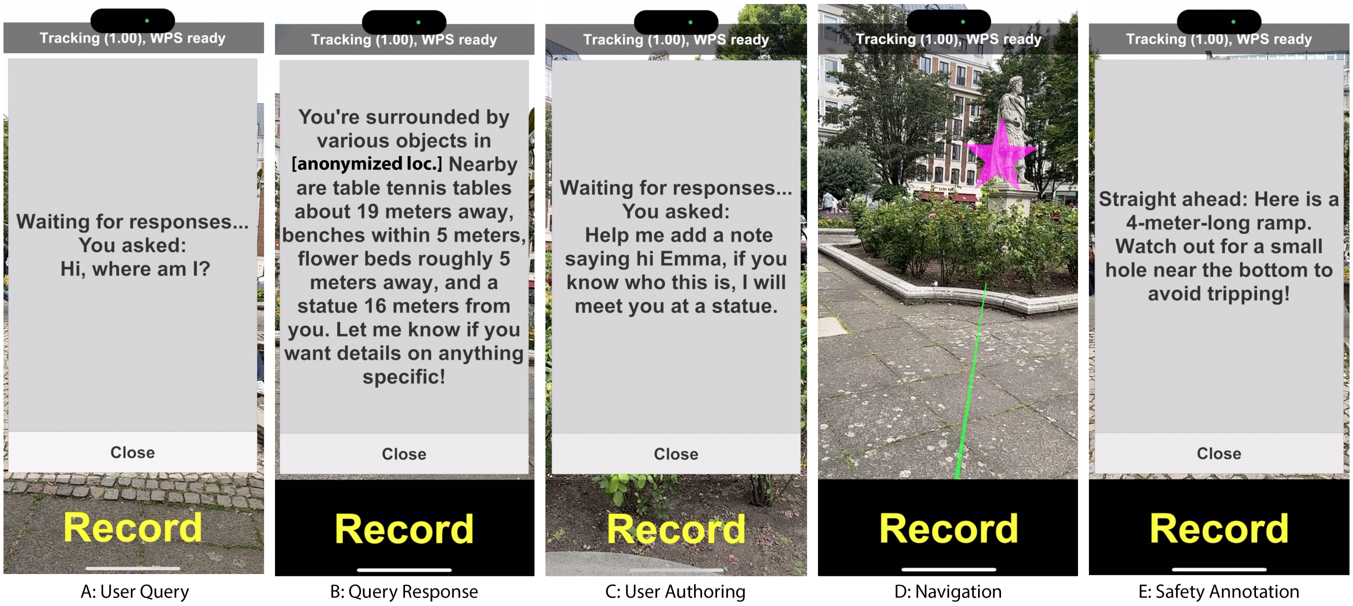

Figure 5: NaviNote's high-contrast visual interface. In (A), the user has asked where they are, and is waiting for a response from the system. In (B), the system responds. In (C), the user asks the system to place an annotation. In (D), the system shows a path to the statue during navigation. In (E), a safety annotation automatically plays. Note that the interface is voice-first, and this visual interface serves to address DG6.

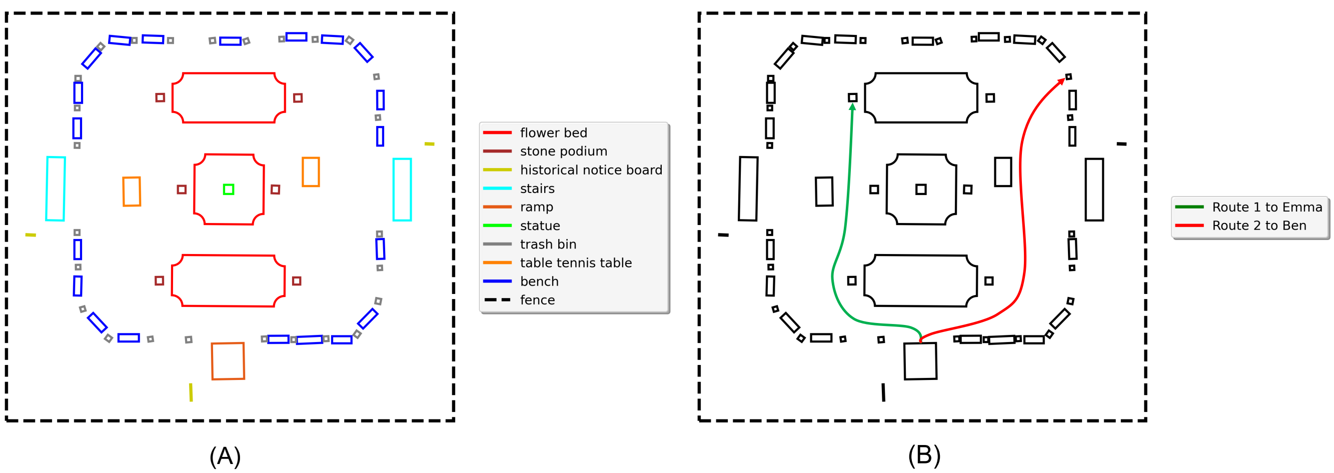

Figure 6: (A) The layout of the local public square for the evaluation study (gates omitted). There are 20 benches, 26 trash bins, two set of stairs, one ramp, six stone podiums, three flower beds, two table tennis tables, three historical notice boards, and one statue. (B) The two routes for the navigation task. Participants start from the ramp and navigate to the furthest podium on the left for "Emma" (green route), and to a trash bin on the right back corner of the square for "Ben" (red route).

to double tap the screen to take a picture and wait 7–10 seconds for the image description [137]. Participants practiced using both systems freely until fully confident about their usage.

5.3.2 Navigation Task. We contextualized the navigation task in the scenario of meeting two friends, named "Emma" and "Ben," at specific locations in the square. We provided initial destination descriptions before the task (e.g., "Emma told you she'd meet you at the podium with the vase, which is the leftmost podium furthest from you if you're starting from the ramp") and marked the destination with large text labels. For NaviNote, an annotation was placed at the two destinations (Supplementary Section ??), and participants could ask the system to provide turn-by-turn instructions to the meeting spots. Although NaviNote supports creating annotations during navigation, participants were asked to explore this feature later during the free-form exploration to ensure consistent comparison of navigation time. For TapTapSee, participants could use it to get scene descriptions as needed during navigation. Researchers followed the participants during the walk to ensure safety, and participants could stop the task at any time. We counterbalanced the order of the two systems (NaviNote or TapTapSee) and the two routes using Latin Square [125], with four participants assigned to each combination randomly.

After each navigation, participants described the route with landmarks they recalled. They also evaluated the system performance using UMUX-LITE [69] on *Perceived Effectiveness* and *Ease of Use*, and NASA Task Load Index (TLX) [49] on *Mental Demand*, *Physical Demand*, *Performance*, and *Frustration*. Participants gave a 7-point Likert scale on each subjective measure. See Supplementary Section **??** for the full script. We also collected participants' qualitative feedback on both systems.

5.3.3 Free-Form Exploration. Participants then conducted a free exploration task within the square using NaviNote. They were contextualized in the scenario where they have some time to explore the square and existing annotations at their own pace before their friends arrived, and were encouraged to create new annotations for audiences of their choosing. During the exploration, they could freely use all features of NaviNote to navigate and query their surroundings. Each participant was required to create at least three annotations.

5.3.4 Exit Interview. We ended our study with a semi-structured interview, discussing annotations participants created, the purposes and target audiences of these annotations, as well as annotations and scenarios they envisioned outside the square (see Supplementary Section **??**). We also asked for feedback on existing annotations, on the design of NaviNote, and suggestions for future in-situ annotation authoring systems.

5.4 Analysis

Since two participants (P17, P18) did not complete the navigation task (Section 5.1), we analyzed the quantitative data from P1–P16 while including qualitative data from all 18 participants. For P1–P16, the navigation task routes and the system order were fully counterbalanced, with four participants for each combination of the route and route order. Our analysis methods are described below.

5.4.1 Measures & Statistical Analysis. In addition to the Likert scale ratings (Section 5.3), we defined two measures for navigation: (1) *Navigation Success*, a binary value indicating whether a participant successfully reached the destination and (2) *Total Landmarks Recalled*, the number of landmarks participants correctly remembered, reflecting their mental map construction [17]. In line with established definition that landmarks should be permanent, unique and identifiable [147], we considered flower bed, statue, stone podium, and table tennis table as landmarks, while excluding transient or repetitive features such as people, benches (20 in the square), and trash bins (26 in the square).

Using the Shapiro-Wilk test, we found none of the measures to be normally distributed. Therefore, we applied a generalized linear mixed model (GLMM) [12] to *Navigation Success* and Aligned Rank Transform (ART) ANOVA on other measures for statistical analysis. For each measure, we included one within-subjects factor, *System*, with two levels: NaviNote and TapTapSee. To validate the counterbalancing, we also included *Order* representing the order of the two systems in the navigation task and found no significant effect of *Order* on any measures. We evaluate statistical significance at $\alpha = .05$. Partial eta-squared (η_p^2) was used to calculate effect size, with 0.01, 0.06, 0.14 representing the thresholds of small, medium and large effects [21, 141].

5.4.2 Qualitative Analysis. The thematic analysis process [14, 20] for this study was identical to that in the Formative Study (Section 3.2), except that the initial open-coding by two researchers was performed over five participants' transcripts (~28% of the data). In addition to participants' transcripts, the research team analyzed all user queries to categorize user questions. We only include questions initiated by participants and exclude those prompted by the research team during the tutorial (Section 5.3). We also analyzed the system responses of all user queries regarding response correctness and types of error.

6 Evaluation Study Results

We report participants' performance in the navigation task and experiences with NaviNote across navigation, exploration, and user query. We then share participants' envisioned future system use cases.

6.1 Navigation Experiences

6.1.1 NaviNote Improved Navigation Performance. We found a significant effect of *System* on *Navigation Success* ($\chi^2(1) = 9.11$, $p = 0.003$), with 14 of 16 participants successfully navigating using NaviNote (P1-P9, P11-P14, P16) compared to only 6 with the baseline, TapTapSee (P4, P5, P7, P13, P15, P16). Of the two participants (P10, P15) who did not complete navigation with NaviNote, P10 stopped about two meters from the destination due to phone tracking drift (Section 7.2.3), and P15 considered a similar nearby object as the destination and stopped following instructions.

6.1.2 Easier Navigation Perceived with NaviNote. We found a significant effect of *System* on *Mental Demand* ($F_{1,15} = 5.36$, $p = 0.035$, $\eta_p^2 = 0.26$), *Self-rated Performance* ($F_{1,15} = 15.71$, $p = 0.001$, $\eta_p^2 = 0.51$) and *Frustration* ($F_{1,15} = 12.00$, $p = 0.003$, $\eta_p^2 = 0.44$), as shown in Figure 7. Compared to navigating using TapTapSee, participants felt that navigating with NaviNote was less mentally demanding ($diff = -6.12$, $p = 0.035$), more successful ($diff = 8.69$, $p = 0.001$), and less frustrating ($diff = -7.38$, $p = 0.004$). Fifteen participants appreciated NaviNote's continuous instructions. Eight participants considered the instructions effective for reducing mental load, as the constant confirmation provided "a lot of support and comfort" (P14). In contrast, fourteen participants felt that TapTapSee was insufficient for navigation, noting that they needed to take "indefinite number of photos" (P10, P14) and only received directions by chance (P6, P10, P14).

While no significant effect was found in *Physical Demand* ($F_{1,15} = 3.14$, $p = 0.097$, $Mean_{\text{NaviNote}} = 1.82$, $Mean_{\text{TapTapSee}} = 2.59$, $SD_{\text{NaviNote}} = 1.42$, $SD_{\text{TapTapSee}} = 2.00$), four participants (P8, P10, P11, P15) reported that taking photos while walking increased their physical demand. P5 and P13 further explained that with TapTapSee, the lack of directional information led them to take smaller steps, which in turn increased the physical load. Nine participants appreciated NaviNote for being hands-free, and eight liked NaviNote for

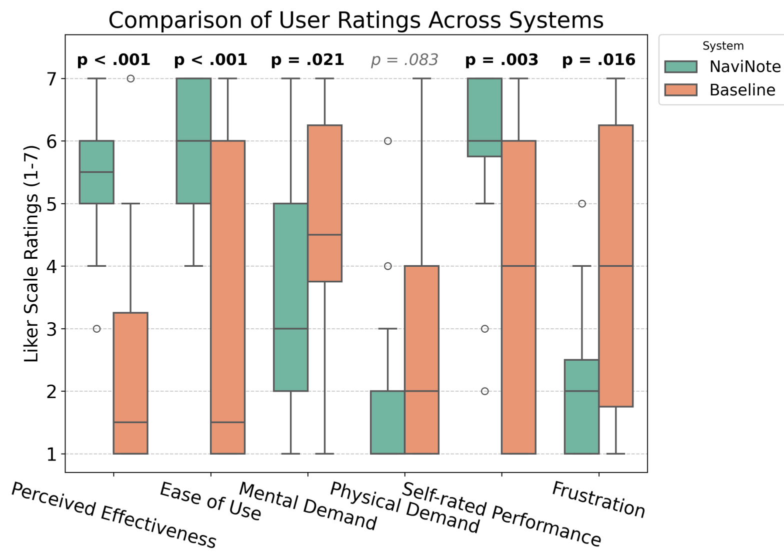

Figure 7: Comparison of users' subjective ratings between NaviNote and the baseline system using UMUX-LITE (Perceived Effectiveness, Ease of Use) and NASA-TLX (Mental Demand, Physical Demand, Self-rated Performance, Frustration)

no need to aim the camera at specific targets. As P13 detailed, "So with TapTapSee, it's expecting you to be able to point the camera at something and being able to find what you're looking at. With [NaviNote], I don't have to look." Compared to TapTapSee, which only had access to the camera frame, NaviNote incorporated the spatial understanding of the entire POI beyond camera view, an essential part of making NaviNote hands-free. Additionally, nine participants disliked TapTapSee for having to stop and perform additional gestures while walking with white canes, describing this process as "completely disorientating" (P7).

6.1.3 Design Considerations for Navigation Cues. We found a significant effect of *System* on *Perceived Effectiveness* ($F_{1,15} = 38.44$, $p < 0.001$, $\eta_p^2 = 0.72$) and *Ease of Use* ($F_{1,15} = 26.00$, $p < 0.001$, $\eta_p^2 = 0.63$), where participants considered NaviNote more effective for navigation ($diff = 12.9$, $p < 0.001$) and easier to use ($diff = 10.2$, $p < 0.001$) than TapTapSee. Participants shared their experience of the audio compass, verbal instructions, and multimodal feedback on safety warnings. They also offered improvement suggestions:

Audio Compass. Eight participants specifically mentioned appreciating the audio compass, describing it as easy to understand (P10, P13) and helpful for finding and confirming the correct direction (6/18). Four participants (P7, P14, P17, P18) further suggested that, in addition to louder beeping signals when facing the correct direction, NaviNote should also provide real-time verbal confirmation. In terms of distraction, P15 found the compass subtle and not distracting, whereas P2 and P16 found it confusing and preferred only verbal corrections.

Verbal Instructions. Eight participants found the turn-by-turn instructions straightforward and easy to follow. Participants exhibited different preferences on the information within the instructions: four (P6, P8, P11, P14) appreciated constant directional assurance, and five liked the occasional distance-to-destination updates, which assured them that they were "getting closer" (P7). However, three participants (P9, P13, P18) found the updates on distance distracting and wanted only immediate direction. P9 and P18 liked the instructions for providing information about "haptic path" (Section 4.2.3). As P9 explained, "If I go anywhere, I try to check if there is a wall or a curve or something to follow to keep in straight line. [... In the experiment I] followed the flower bed, which was a good guideline."

Four participants (P7–P9, P12) noted a learning curve for the instructions. Five participants wanted NaviNote to explicitly say when they were going the wrong way, rather than just telling them to "turn backward." They also suggested using more natural phrases, such as "turn around" instead of "backward" (5/18), and simplifying egocentric directions to four basic options (i.e., left, right, forward, backward) rather than eight (P8, P9).

Multimodal Feedback. NaviNote provided haptic vibrations when playing safety annotations (Section 4.2.4). P13 appreciated the multimodal feedback as it made the system more robust: "I wasn't just relying on one thing. If something stopped working but I felt a vibration, I still knew there was something happening or there was something to be aware of."

6.2 Annotation Experiences

6.2.1 NaviNote Helps Users Understand their Surroundings. We found a significant effect of *System* on *Total Landmarks Recalled* ($F_{1,15} = 15.21$, $p = 0.001$, $\eta_p^2 = 0.50$), with participants remembering more landmarks with NaviNote than TapTapSee ($diff = 8.75$, $p = 0.001$). Thirteen participants mentioned how the annotations in NaviNote helped them learn about their surroundings, e.g., by providing general descriptions of the square and by offering specific information tailored to their needs. Nine participants praised the annotations for giving a general impression of the layout and available amenities. As P15 recounted, "[with NaviNote] I learned there are branches, there are benches, there are statues, there are flowers, and people sitting around." P10 further noted that crowdsourced annotations in NaviNote could help identify "what services or amenities are available" and could help him plan for social activities.

In terms of specific information, four participants (P1, P4, P7, P14) noted appreciating historical annotations about the square and the statue. Three participants (P5, P13, P18) valued information about local attractions, such as descriptions of flowers and their species (P13, P18), which raised their awareness of features they would otherwise have difficulty seeing (P5, P9). Notably, four participants (P12, P15, P17, P18) praised NaviNote for providing color information about objects. P12 noted that color information about obstacles with low contrast against the background (e.g., a gray raised flower bed vs. gray ground) made them easier to detect and avoid.

6.2.2 Experiences with Different Types of Annotations. In addition to agreeing that the layout and amenities annotations helped them understand their surroundings (Section 6.2.1), participants highlighted the help of safety and accessibility annotations in supporting safer mobility. Eleven participants noted appreciating these annotations during navigation and free exploration, as they provided advanced alerts of obstacles, helping them avoid them. As P14 described, "[with safety annotations] it was good to know even before I get there that I'm going to encounter a raised [flower] bed." Three participants (P3, P5, P10) noted liking NaviNote for *automatically* playing safety alerts. Four participants (P1, P7, P10, P17) further emphasized the usefulness of warnings about overhanging obstacles (e.g., tree branches), which conventional aids, like white canes, typically miss. As P1 explained, "[NaviNote] would be a complementary assistance to white cane... because the idea is to avoid bumping my head into a barrier."

Participants expressed diverse opinions on attraction and experience annotations. Several participants (6/18) appreciated attraction annotations for sparking interest in nearby areas rather than "easily missing them" (P18), and three (P4, P8, P17) even considered them their favorite type of annotation. In contrast, four participants (P3, P5, P6, P10) chose to filter out attraction annotations due to lack of interest or usefulness. Notably, all four of these participants were blind. Opinions on experience annotations were similarly mixed: three participants (P1, P8, P16) liked how they shared personal experiences of the area, especially from BLV users' perspectives (P8), while P10 preferred practical information over subjective experiences, and P5 noted only being interested in such annotations from friends.

6.2.3 Annotation Authoring. All participants created annotations successfully. Building on existing annotations and querying NaviNote for additional information, they actively created new annotations to share with themselves as reminders (4/18), with friends and family (7/18), with the BLV community (8/18), and with the public (7/18). In addition, participants (12/18) appreciated the ability to edit and delete annotations, as this allowed them to correct mistakes (P8, P12, P17) and prevent AI misunderstandings (P1, P10). Most annotations that participants created fit within our existing annotation taxonomy: safety (6/18), accessibility (P7), amenity (7/18), layout (5/18), attraction (P2, P8, P9), and experience (8/18). Interestingly, three participants (P4, P13, P14) created annotations that aimed to *request information from other users*. For instance, when feeling unsure about the identity of a statue, P14 created an annotation asking other users to "find it out" and tell her. This type of annotation fell outside of our original categorization, but would nonetheless provide valuable opportunities for users to communicate with each other; thus, we add a final category to our taxonomy, "Request."

6.2.4 Control Over Annotation Access. Six participants appreciated that NaviNote allowed them to filter annotations by type. Seven participants suggested refinements to these "information layers" with respect to annotation category, content, and context. P7 proposed finer-grained annotation categorization, while both P3 and P7 recommended filtering by content rather than type to retrieve useful information. For non-safety annotations, P10 wanted annotations linked to places of personal significance to automatically play, P14 suggested applying location-aware filtering (e.g., outside the square vs. in the square), P13 and P16 recommended merging duplicate information (e.g., flowers in two identical flower beds), and P11 suggested accessing these annotations only in first-time visits.

While participants valued the different access modes for annotations (i.e., automatic, prompted, or silent), P13 found asking the system for the prompted annotations to be tedious. She suggested incorporating lightweight gestures for annotation access, such as using the volume button to quickly play or dismiss prompted annotations.

6.3 User Query Experiences

We evaluate the performance of NaviNote's voice-based question answering regarding the accuracy of NaviNote and participants' interaction experiences. We further analyze the types of user query and NaviNote's response time.

6.3.1 Accuracy. Participants asked a total of 605 questions. Among these questions, 501 (82.8%) were correctly answered, 34 (5.6%) were incorrect, 13 (2.2%) were partially correct, 39 (6.5%) were misrecognized by the speech-to-text model (e.g., due to accents, background noise or low volume), and the AI failed to answer the remaining 18 questions (3.0%) for various reasons listed below.

Half of the incorrect answers (17/34, 50%) were caused by failing to retrieve relevant information for generating answers or invoking actions. These failures stemmed from miscommunication between the orchestrator and the agents, such as not passing or returning the correct parameters (e.g., when a user asked about the statue's history, the orchestrator only instructed the annotation agent to search for history-related annotations without specifying statue-related ones). The second most common cause was the orchestrator invoking the wrong agent as it misunderstood the user query due to two similar terms (e.g., "attractions" vs. "attraction annotations," which prompted the orchestrator to invoke the internet search agent or annotation agent respectively; 6/34, 17.6%). Other reasons included the AI failing to retrieve chat memory for context-based answers (5/34, 14.7%), retrieving incorrect information (e.g., using users' Unity coordinates instead of real-world positions; 5/34, 14.7%), and AI hallucinations (1/34, 3.0%) For the 13 partially correct answers, the most frequent cause was the orchestrator generating the correct answer without invoking the appropriate agent to trigger the corresponding frontend action (e.g., start navigation; 8/13, 61.5%).

For the 18 questions that NaviNote failed to answer, eleven (61.1%) occurred because the system lacked available information (e.g., the participant asked "How is the weather today?" while NaviNote has no access to the date). Six cases (33.3%) were due to "Too Many Requests" errors from OpenAI API, and one case was due to OpenAI incorrectly filtering safe content.

6.3.2 Conversational Voice-based Interaction Experiences. Thirteen participants explicitly appreciated NaviNote for the conversational and voice-based interaction design, describing it as "intuitive and user-friendly" (P9). P7 also said, "I've never seen a system where you can say instructions. You [usually] have to go into menus to do things." P17 further suggested that, in addition to voice control, the system should incorporate a graphical user interface to allow exploring available functions with screen readers.

We also observed that 11 out of the 18 participants interacted with the system in a conversational style. For example, when creating an annotation, instead of immediately providing the content, participants would first tell the system they wanted to create a annotation, wait for the system to ask for content, then provide it in a follow-up query. Five participants also asked NaviNote to repeat responses.

6.3.3 Question Types. As participants often asked similar questions repeatedly, we counted the number of participants who asked each type of question rather than the total occurrences, to prevent any single participant from disproportionately affecting the distribution of question types. Among all questions initiated by the participants, the most common three types were directly related to system features: (1) All 18 participants inquired about annotation content and asked to author annotations; (2) Seventeen participants asked to customize NaviNote; and (3) Twelve participants asked about navigating to specific locations. In addition to questions related to system features, participants asked four types of questions related to their surroundings. Ten participants (55.6%) asked for **general descriptions of their surroundings**, such as the layout and size of the square. Nine participants inquired about the **objects and activities at specific locations**, such as number of benches and available services at the square. Nine participants asked about **object details** in or beyond the camera view. For example, P17 asked, "Could you describe these birds for me?" after hearing bird sounds. Eight participants asked about **object locations relative to themselves** (e.g., "Is the flower bed [that] the one attraction note [talks about] on my left or right?").

6.3.4 Response Time. NaviNote answered user queries in an average of 10.8 seconds (SD = 8.3, Max = 41.0, Min = 1.2, see Figure 8). The variation was partly due to question complexity (e.g., "Hello!" vs. "My friend Emma left me a note. Can you help me find where she is?"), while longer delays were typically caused by unstable internet or occasional "Too Many Requests" errors from the OpenAI API, upon which the API waited for around 20 seconds before retrying. Six participants (P2, P5, P6, P9, P17, P18) expressed dissatisfaction (*Mean response time* = 15.1*s*), noting that the system should respond more quickly, while P15 (*Mean response time* = 10.8*s*) was satisfied and described it as "responding very quickly."

6.4 Independence through Annotations: When Are Annotations Useful?

Eight participants highlighted that NaviNote gave them **more independence** compared to asking others for help. As P13 explained, "I don't want to always have to wait for somebody sighted. I quite like to get out, but people always get worried." Participants appreciated that NaviNote enabled them to take initiative rather than receive information passively (5/18), by asking follow-up questions and actively leaving notes. Most participants (14/18) noted that they would feel comfortable exploring alone with NaviNote in the future.

Participants also envisioned potential use cases for NaviNote beyond the experiment. These cases generally fell into four categories: navigation, relaxation and exploration, information gathering, and social interactions.

6.4.1 Navigation. Several participants (8/18) envisioned using annotations to mark interested spots and navigate back in later days. For example, five participants wanted to mark indoor public spaces, such as airport gates, train stations, and malls for easier access. P7 also proposed returning home with NaviNote: "My biggest fear is getting stuck in the house, not able to come out because I've always got this fear: The minute I leave my house, I won't be able to find my house again because all the houses look the same and feel the same from outside."

6.4.2 Relaxation and Exploration. Participants wanted to use NaviNote for museum exploration, street wandering, and travel (8/18) to help independent exploration, which is often challenging. For example, P5 said, "Usually I would just walk [past exhibits]. That's why I hate going to museums. There's a load of stuff I can't see." With NaviNote, however, she could discover visual attractions that she normally wouldn't see.

6.4.3 Information Gathering. Participants saw potential value in viewing others' comments in advance. As P4 envisioned, "I can see a map of [annotations] and then there's like little stars or something. I click it and I can hear it. I would actually do that." They also wanted to gather information about tourist attractions (P5), or leave notes to request information, so that other BLV individuals visiting the same location later could benefit (P13, P14).

6.4.4 Social Interactions. Participants envisioned leaving audio notes for social purposes (6/18), such as marking meeting locations or sharing information with friends. For example, P13 imagined a scenario where, if she were running late, her friends could leave her an annotation about a new meeting spot. P16 described annotations as a feature that "makes [the environment] more interactive" by sharing information with other users, and P4 further emphasized that accessible annotations can promote social interaction: "Can you imagine that? You can actually interact with people. Because a lot of people, they don't interact that much... So I like the audio note element [for allowing more interactions]."

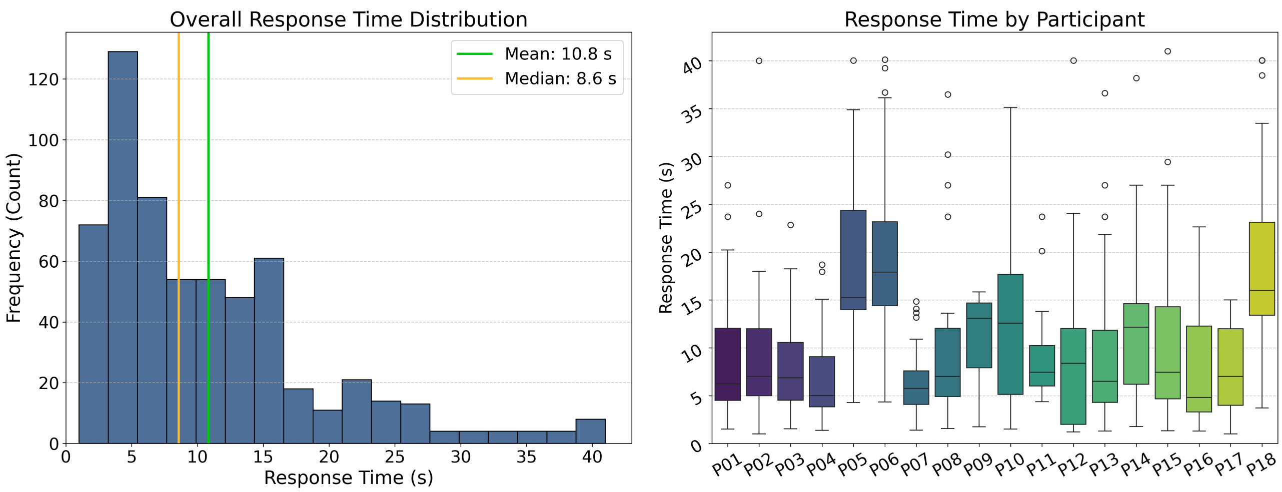

Figure 8: Left: Distribution of NaviNote's response time per user query, with mean response time to be 10.8 seconds and median response time 8.6 seconds. Most queries were answered within 15 seconds. Right: Response times for individual participants, with variation between participants primarily due to unstable internet connections in the field.

7 Discussion

Our evaluation showed that NaviNote effectively supports BLV participants in independently authoring annotations, navigating the last few meters to POIs, and exploring local public places. Participants also envisioned using it across various use-cases. In this section, we discuss design challenges we discovered related to spatial annotation authoring tools, reflect on agent-based accessibility systems, and outline limitations and future directions. Note that in this section we discuss broad topics, while we outline specific design implications for NaviNote in Section 6.

7.1 Design Challenges for Spatial Annotation Authoring Tools

Participants praised NaviNote for supporting scene exploration and enabling them to author and access spatial annotations. They also suggested ways to improve retrieval mechanisms and safety, as discussed below.

7.1.1 How can annotations maintain scene alignment over time? While crowdsourced annotations anchored to spatial locations can help BLV users explore their surroundings, they also introduce the challenge of staying physically aligned after changes in the scene. During free exploration, participants were concerned that objects of interest may be moved or removed, making associated annotations obsolete. In addition to manual edits via crowdsoucing, advances in 3D change detection [80, 81, 119] opens up the possibility of automatic detection and realignment of annotations [106]. We suggest that future systems explore automatically detecting environmental changes and realigning or removing outdated annotations to ensure they remain reliable.

7.1.2 How can annotations be kept true, safe, and ethical? As with many user-generated content (UGC) platforms, crowdsourcing introduces risks of rumors and malicious content, which can undermine users' experiences and even threaten safety [18, 131, 132], especially for BLV users. Existing mechanisms from UGC platforms can be adapted to moderate such annotations: (1) "Official/Verified" annotations from trusted users, similar to the "blue checkmark" system on X [146] or Instagram [54]; (2) Upvote/downvote mechanisms, where heavily downvoted annotations are removed [66]; and (3) AI filtering to automatically detect and remove unsafe notes. That said, AI-based approaches are not always reliable due to inaccuracies and biases in training data [4, 11]. For instance, in our study, AI incorrectly filtered out a harmless note about a nearby mosque (1/605). Future work should explore strategies to ensure spatial annotations remain genuine, safe, and ethical.

7.1.3 How to manage annotation access and prevent information overload? In the current prototype of NaviNote, all notes were made publicly accessible to simplify evaluation. However, our findings show that participants frequently authored annotations intended only for themselves, their acquaintances, or specific BLV audiences (Section 6.2.3), which further supports the social annotation needs reported in FootNotes [36]. To better protect end-user privacy, future systems should incorporate fine-grained access control mechanisms that verify the intended audience before an annotation is shared. Moreover, systems should offer options for anonymity and warn users when they disclose sensitive or identifiable information [63]. For example, an annotation such as "Meet me here" should only be visible to trusted contacts, and a note like "This is my favorite cafe in my neighborhood" should be shareable anonymously.

In addition to privacy considerations, filtering information based on target groups can help prevent information overload. NaviNote currently prioritizes safety- and accessibility-related annotations and skips notes that cannot be played within 15 seconds after a user approaches them (Section 4.2.4). Future systems should further prioritize annotations based on both the identity of the contributor (e.g., assigning higher priority to notes from known or trusted people) and community ratings within the target group (e.g., upvotes or verification mechanisms as discussed in Section 7.1.2). These strategies can help ensure that users receive relevant, trustworthy, and manageable information.

7.2 Usability Challenges of Exploration Systems in the Wild

While NaviNote provided substantial support in last-few-meters navigation and free exploration, our evaluation also revealed several usability challenges, which we reflect on below.

7.2.1 Crowdsourced Scanning. NaviNote is built on the assumption that scans can be crowdsourced from sighted volunteers and friends, a process that has been used to capture locations world-wide [13, 99, 107], and can be completed online [101] and offline [100]. With consent, BLV user sessions could also be used to keep scans up-to-date. In the future, we also envision gathering scans from drones [96, 150], delivery robots [82], and automated vehicles [8]. Given VPS scans still have limited coverage [13], in the meantime, research should investigate methods to provide seamless transitions between more/less accurate localization methods, e.g., between VPS and GPS, such that users can have *at least some* guidance no matter where they are. Furthermore, research should explore mechanisms to automatically update scene understanding, e.g., whenever users localize and contribute new scans (Section 7.1.1).

7.2.2 Guidance for Independent Localization. Despite not needing to aim NaviNote at specific targets during use, NaviNote's VPS still requires an initial localization, in which users sweep the camera back and forth until the system identifies its location [38, 104]. Although we did not formally test this localization process with every user, as it was out of scope for our study, ten participants tested independent localization with eight succeeding within one minute (80%). Future work should explore how to improve localization for BLV users to enhance the generalizability and robustness of our findings, e.g., by designing effective audio and haptic guidance for this process similar to [19, 85, 121, 139].

7.2.3 Detecting and Notifying Users of Drift. While VPS enabled high accuracy and continuous localization, some participants still experienced mid-use drift, where the phone failed to track its 3D position accurately, causing the navigation algorithm to misguide them (e.g., P10 stopped about two meters from the destination). To accurately reflect NaviNote's performance, we did not intervene when drift occurred. Even though participants were instructed before the task to navigate as they normally do (e.g., with white canes) and only treat NaviNote's instructions as an additional reference point, they still exhibited a high-level of trust in the system during the experiment. Future tools should explore mechanisms to counter mid-use drift (e.g., visual-based pose estimation [71] and localization [2], reprojection [139]), as well as using MLLMs for visual-semantic reasoning to correct AR content [106]. Moreover, future research should explore effective ways to detect and alert users of drift, such as reporting localization confidence and instructing users to re-localize. Annotating the environment with landmark-specific cues (e.g., haptic textures, smells, or characteristic sounds [76, 127]) may also improve landmark recognition and further support users' self-localization when drift occurs.

7.3 AI Agent-based Systems for Accessibility

Most of our BLV participants expressed trust in NaviNote. As P8 noted, "The way it gives directions is very clear. And then that's how it builds up trust." However, there were one case (1/605) where the system generated incorrect answers and the participant accepted fabricated information (although this was rare). For example, P2 believed the statue depicted someone other than who it was based on a system hallucination. Future systems should make users aware of AI's limitations (e.g., the types of questions it can reliably answer) and the possibility of hallucinated responses before use [51, 59, 94, 95]. They should also investigate how to provide richer context and details to help BLV users recognize errors independently [51], such as including information sources and image descriptions in the response (e.g., "This information came from the internet" vs. "... from user X's annotation"). More importantly, future work should focus on building more reliable and explainable AI systems [6, 27, 28].

AI systems can also suffer from response latency [67, 151]. In our study, some participants expressed wanting faster responses (Section 6.3). Since NaviNote incorporated the "Orchestrator" pattern [110], response time was influenced by both the base model latency [1] and question complexity, which could require one or multiple rounds of communication between the "Orchestrator" and specialized agents. Beyond model-level improvements, future research should also investigate speed optimization of agent-based architectures, while tailoring responses to BLV users' needs [42].

7.4 Limitations and Future Directions

This work has several limitations and future directions, in addition to those mentioned in the discussion above. First, although the preset annotations used in our system evaluation were derived from participants' desired annotations identified in our formative study, they were tested in single-user sessions. As a result, our evaluation may not fully capture the dynamics of community-based, crowdsourced annotations, e.g., how BLV users might interpret colloquial or incomplete annotations. Moreover, we evaluated NaviNote in a local public square with comparatively few objects and relatively simple routes, which may not fully reflect BLV users' actual daily settings. Additionally, further effects and concerns may have eluded us based on our relatively small sample size of 16 users. In summary, future research should investigate BLV users' experiences with crowdsourced annotations in longitudinal and multi-user studies, test such systems in more complex and realistic public environments, and include more participants with diverse visual conditions.

8 Conclusion

Thanks to a formative study with 24 blind and low vision (BLV) participants, we contribute the design and implementation of NaviNote. NaviNote is an accessible voice-based interface, incorporating Visual Positioning Systems and Multimodal Large Language Models to enable BLV users to author in-situ annotations and navigate effectively in last-few-meters scenarios. Through an evaluation with 18 BLV participants, we found that NaviNote significantly improved BLV users' navigation performance, helped them better understand their surroundings, and enabled them to actively annotate the environment. We also discuss design considerations for future accessible annotation systems and provide an initial taxonomy of BLV users' desired spatial annotations to guide future crowdsourcing efforts.

Acknowledgments

We thank Bryan Tan, Sara Gomes Vicente, Michael Firman, Jakub Powierza, Chris Shaw, Michael Miller, Ben Benfold, Alyssa Coley, Effie Shum, and Belinda Manfo from Niantic Spatial, Inc. for their assistance in the successful execution of both studies and in system deployment. We thank Beyond Sightloss for facilitating participant recruitment in both studies.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Yongbao Ai, Ting Rui, Xiaoqing Yang, Jiantao He, Fu Lei, Jianbin Li, and Lu Ming. 2020. Visual SLAM in dynamic environments based on object detection. *Defence Technology* (2020). https://doi.org/10.1016/J.DT.2020.09.012

[3] Rahaf Alharbi, Pa Lor, Jaylin Herskovitz, Sarita Schoenebeck, and Robin N Brewer. 2024. Misfitting With AI: How Blind People Verify and Contest AI Errors. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–17.

[4] Oghenefejiro Isaacs Anigboro, Charlie M Crawford, Grace Proebsting, Danaë Metaxa, and Sorelle A Friedler. 2024. Identity-related Speech Suppression in Generative AI Content Moderation. *arXiv preprint arXiv:2409.13725* (2024).

[5] Apple Inc. 2025. Add a bibliography in Pages on Mac. https://support.apple.com/guide/pages/bibliography-tan0e6bf8310/mac Accessed: 2025-08-29.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[7] Takumi Asakura. 2021. Bone conduction auditory navigation device for blind people. *Applied Sciences* 11, 8 (2021), 3356.

[8] Daria Belkouri, Richard Laing, and David Gray. 2022. Through the eyes of Autonomous Vehicles–using laser scanning technology to engage the public via the analysis of journeys seen from a different perspective. *Transportation Research Procedia* 60 (2022), 496–503.

[9] Frank Bentley and Santosh Basapur. 2012. StoryPlace. Me: the path from studying elder communication to a public location-based video service. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. 777–792.

[10] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 65–72.

[11] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*. Springer, 405–415.

[12] Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24, 3 (2009), 127–135.

[13] Eric Brachmann and Victor Adrian Prisacariu. 2025. Building a Large Geospatial Model to Achieve Spatial Intelligence. https://nianticlabs.com/news/largegeospatialmodel.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[15] Robin N. Brewer. 2018. Facilitating discussion and shared meaning: Rethinking co-design sessions with people with vision impairments. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare* (New York, NY, USA) *(PervasiveHealth '18)*. Association for Computing Machinery, New York, NY, USA, 258–262. https://doi.org/10.1145/3240925.3240981

[16] 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*. 1–18.

[17] Ruijia Chen, Junru Jiang, Pragati Maheshwary, Brianna R Cochran, and Yuhang Zhao. 2025. VisiMark: Characterizing and Augmenting Landmarks for People with Low Vision in Augmented Reality to Support Indoor Navigation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[18] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the international aaai conference on web and social media*, Vol. 9. 61–70.

[19] SeungA Chung, Kyungyeon Lee, and Uran Oh. 2021. Understanding the two-step nonvisual omnidirectional guidance for target acquisition in 3d spaces. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 339–346.

[20] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.

[21] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

[22] Vlad Coroama and Felix Röthenbacher. 2003. The Chatty Environment–providing everyday independence to the visually impaired. In *Workshop on ubiquitous computing for pervasive healthcare applications at UbiComp*.

[23] James M Coughlan, Brandon Biggs, Marc-Aurèle Rivière, and Huiying Shen. 2020. An audio-based 3d spatial guidance ar system for blind users. In *International Conference on Computers Helping People with Special Needs*. Springer, 475–484.

[24] James M Coughlan and Joshua Miele. 2017. AR4VI: AR as an accessibility tool for people with visual impairments. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, 288–292.

[25] James M Coughlan, Huiying Shen, and Brandon Biggs. 2020. Towards accessible audio labeling of 3D objects. In *Journal on technology and persons with disabilities:... Annual International Technology and Persons with Disabilities Conference*, Vol. 8. 210.

[26] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200.

[27] David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. 2024. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624* (2024).

[28] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys* 55, 9 (2023), 1–33.

[29] Ian Fette and Alexey Melnikov. 2011. *The websocket protocol*. Technical Report.

[30] Alexander Fiannaca, Ilias Apostolopoulous, and Eelke Folmer. 2014. Headlock: a wearable navigation aid that helps blind cane users traverse large open spaces. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 19–26.

[31] J. Finlay, Michael H. Esposito, Min Hee Kim, Iris N. Gomez-Lopez, and P. Clarke. 2019. Closure of 'third places'? Exploring potential consequences for collective health and wellbeing. *Health & place* 60 (2019), 102225. https://doi.org/10.1016/j.healthplace.2019.102225

[32] American Foundation for the Blind. 2009. GPS Made Simple: A Review of the Trekker Breeze. https://afb.org/aw/10/2/16127. [Accessed 09-11-2025].

[33] Jon E Froehlich, Alexander Fiannaca, Nimer Jaber, Victor Tsara, and Shaun Kane. 2025. Streetviewai: Making street view accessible using context-aware multimodal ai. *arXiv preprint arXiv:2508.08524* (2025).

[34] Bhanuka Gamage, Thanh-Toan Do, Nicholas Seow Chiang Price, Arthur Lowery, and Kim Marriott. 2023. What do blind and low-vision people really want from assistive smart devices? Comparison of the literature with a focus study. In *Proceedings of the 25th international ACM SIGACCESS conference on computers and accessibility*. 1–21.

[35] Aura Ganz, James M Schafer, Yang Tao, Carole Wilson, and Meg Robertson. 2014. PERCEPT-II: Smartphone based indoor navigation system for the blind. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society*. IEEE, 3662–3665.

[36] Cole Gleason, Alexander J Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2018. FootNotes: Geo-referenced audio annotations for nonvisual exploration. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[37] Cole Gleason, Anhong Guo, Gierad Laput, Kris Kitani, and Jeffrey P Bigham. 2016. VizMap: Accessible visual information through crowdsourced map reconstruction. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. 273–274.

[38] Google. 2025. Build global-scale, immersive, location-based AR experiences with the ARCore Geospatial API. https://developers.google.com/ar/develop/geospatial.

[39] Google. 2025. *Google Maps*. https://www.google.com/maps

[40] William G Griswold, Patricia Shanahan, Steven W Brown, Robert Boyer, Matt Ratto, R Benjamin Shapiro, and Tan Minh Truong. 2004. ActiveCampus: experiments in community-oriented ubiquitous computing. *Computer* 37, 10 (2004),

73–81.
[41] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. 2023. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. arXiv:2309.16650 [cs.RO] https://arxiv.org/abs/2309.16650
[42] Antonio Gulli, Lavi Nigam, Julia Wiesinger, Vladimir Vuskovic, Irina Sigler, Ivan Nardini, Nicolas Stroppa, Sokratis Kartakis, Narek Saribekyan, and Alan Bount. 2025. Agents companion. *Kaggle/Google, Tech. Rep., February* (2025).
[43] Anhong Guo. 2018. Crowd-AI Systems for Non-Visual Information Access in the Real World. In *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 216–219.
[44] Yu Hao, Alexey Magay, Hao Huang, Shuaihang Yuan, Congcong Wen, and Yi Fang. 2024. ChatMap: A Wearable Platform Based on the Multi-modal Foundation Model to Augment Spatial Cognition for People with Blindness and Low Vision. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 129–134.
[45] Yu Hao, Fan Yang, Hao Huang, Shuaihang Yuan, Sundeep Rangan, John-Ross Rizzo, Yao Wang, and Yi Fang. 2024. A multi-modal foundation model to assist people with blindness and low vision in environmental interaction. *Journal of Imaging* 10, 5 (2024), 103.
[46] 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 Transactions on Accessible Computing (TACCESS)* 6, 2 (2015), 1–23.
[47] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 631–640.
[48] Peter E Hart, Nils J Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107.
[49] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
[50] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2023. Hacking, switching, combining: understanding and supporting DIY assistive technology design by blind people. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[51] Jonggi Hong and Hernisa Kacorri. 2024. Understanding How Blind Users Handle Object Recognition Errors: Strategies and Challenges. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–15.
[52] Viktor Horvath, Charles Toth, and Arpad Barsi. 2025. Investigating the accuracy of Google's Visual Positioning System. In *2025 IEEE/ION Position, Location and Navigation Symposium (PLANS)*. 1594–1600. https://doi.org/10.1109/PLANS61210.2025.11028337
[53] Jonathan Huang, Max Kinateder, Matt J Dunn, Wojciech Jarosz, Xing-Dong Yang, and Emily A Cooper. 2019. An augmented reality sign-reading assistant for users with reduced vision. *PloS one* 14, 1 (2019), e0210630.
[54] Instagram Help Center. [n. d.]. Get Verified on Instagram. https://help.instagram.com/733907830039577. Accessed: 2025-09-03.
[55] Toru Ishikawa, Hiromichi Fujiwara, Osamu Imai, and Atsuyuki Okabe. 2008. Wayfinding with a GPS-based mobile navigation system: A comparison with maps and direct experience. *Journal of environmental psychology* 28, 1 (2008), 74–82.
[56] Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. 2024. Identifying Crucial Objects in Blind and Low-Vision Individuals' Navigation. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.
[57] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. 2011. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. 203–210.
[58] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1827–1843.
[59] 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*. 1624–1638.
[60] Yuka Kaniwa, Masaki Kuribayashi, Seita Kayukawa, Daisuke Sato, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. 2024. ChitChatGuide: Conversational Interaction Using Large Language Models for Assisting People with Visual Impairments to Explore a Shopping Mall. *Proceedings of the ACM on Human-Computer Interaction* 8, MHCI (2024), 1–25.
[61] Brian FG Katz, Slim Kammoun, Gaëtan Parseihian, Olivier Gutierrez, Adrien Brilhault, Malika Auvray, Philippe Truillet, Michel Denis, Simon Thorpe, and Christophe Jouffrais. 2012. NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality* 16, 4 (2012), 253–269.
[62] Brian FG Katz and Lorenzo Picinali. 2011. Spatial audio applied to research with the blind. *Advances in sound localization* 1991 (2011), 225–250.
[63] Kat Krol and Sören Preibusch. 2016. Control versus effort in privacy warnings for webforms. In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*. 13–23.
[64] Masaki Kuribayashi, Tatsuya Ishihara, Daisuke Sato, Jayakorn Vongkulbhisal, Karnik Ram, Seita Kayukawa, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2023. Pathfinder: Designing a map-less navigation system for blind people in unfamiliar buildings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[65] Masaki Kuribayashi, Kohei Uehara, Allan Wang, Shigeo Morishima, and Chieko Asakawa. 2025. Wanderguide: Indoor map-less robotic guide for exploration by blind people. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.
[66] Clifford A Lampe. 2006. *Ratings use in an online discussion system: The Slashdot case*. University of Michigan.
[67] Jaewook Lee, Filippo Aleotti, Diego Mazala, Guillermo Garcia-Hernando, Sara Vicente, Oliver James Johnston, Isabel Kraus-Liang, Jakub Powierza, Donghoon Shin, Jon E. Froehlich, Gabriel Brostow, and Jessica Van Brummelen. 2025. ImaginateAR: AI-Assisted In-Situ Authoring in Augmented Reality. arXiv:2504.21360 [cs.HC] https://arxiv.org/abs/2504.21360
[68] Max Jwo Lem Lee, Li-Ta Hsu, Hoi-Fung Ng, and Shang Lee. 2020. Semantic-Based VPS for Smartphone Localization in Challenging Urban Environments. *arXiv preprint arXiv:2011.10743* (2020).
[69] James R Lewis, Brian S Utesch, and Deborah E Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2099–2102.
[70] Chu Li, Katrina Oi Yau Ma, Michael Saugstad, Kie Fujii, Molly Delaney, Yochai Eisenberg, Delphine Labbé, Judy L Shanley, Devon Snyder, Florian PP Thomas, et al. 2024. "I never realized sidewalks were a big deal": A Case Study of a Community-Driven Sidewalk Accessibility Assessment using Project Sidewalk. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
[71] Peiliang Li, Tong Qin, Botao Hu, Fengyuan Zhu, and S. Shen. 2017. Monocular Visual-Inertial State Estimation for Mobile Augmented Reality. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2017), 11–21. https://doi.org/10.1109/ISMAR.2017.18
[72] Haozhe Lin, Jiangtao Gong, Yu Wang, Jinsong Zhang, Bing Bai, Yan Zhang, Luyao Wang, Chenyu Wei, Yancheng Cao, Kun Li, et al. 2025. AI system facilitates people with blindness and low vision in interpreting and experiencing unfamiliar environments. *npj Artificial Intelligence* 1, 1 (2025), 7.
[73] Jiawei Liu, Hao Tang, WH Seiple, and Zhigang Zhu. 2022. Annotating storefront accessibility data using crowdsourcing. In *Journal on Technology & Persons with Disabilities, Scientific/Research Proceedings of 2022 CSUN Assistive Technology Conference*, Vol. 10.
[74] Sendero Group LLC and HumanWare Group. 2015. Sendero GPS BrailleNote BT and VoiceNote BT. https://senderogroup.com/support/btgpsv2015manual.htm. [Accessed 18-11-2025].
[75] Jacobus Cornelius Lock, Grzegorz Cielniak, Nicola Bellotto, et al. 2017. A Portable Navigation System with an Adaptive Multimodal Interface for the Blind.. In *AAAI Spring Symposia*.
[76] Jack M Loomis, Roberta L Klatzky, Reginald G Golledge, et al. 2001. Navigating without vision: basic and applied research. *Optometry and vision science* 78, 5 (2001), 282–289.
[77] Chen-Lung Lu, Zi-Yan Liu, Jui-Te Huang, Ching-I Huang, Bo-Hui Wang, Yi Chen, Nien-Hsin Wu, Hsueh-Cheng Wang, Laura Giarré, and Pei-Yi Kuo. 2021. Assistive navigation using deep reinforcement learning guiding robot with UWB/voice beacons and semantic feedbacks for blind and visually impaired people. *Frontiers in Robotics and AI* 8 (2021), 654132.
[78] Changsheng Lv, Mengshi Qi, Xia Li, Zhengyuan Yang, and Huadong Ma. 2023. SGFormer: Semantic Graph Transformer for Point Cloud-Based 3D Scene Graph Generation. (2023), 4035–4043. https://doi.org/10.1609/aaai.v38i5.28197
[79] Alexey Magay, Dhurba Tripathi, Yu Hao, and Yi Fang. 2024. A Light and Smart Wearable Platform with Multimodal Foundation Model for Enhanced Spatial Reasoning in People with Blindness and Low Vision. In *European Conference on Computer Vision*. Springer, 323–339.
[80] Murari Mandal, Vansh Dhar, Abhishek Mishra, and Santosh Kumar Vipparthi. 2019. 3DFR: A Swift 3D Feature Reductionist Framework for Scene Independent Change Detection. *IEEE Signal Processing Letters* 26 (2019), 1882–1886. https://doi.org/10.1109/LSP.2019.2952253
[81] Murari Mandal, Vansh Dhar, Abhishek Mishra, Santosh Kumar Vipparthi, and M. Abdel-Mottaleb. 2020. 3DCD: Scene Independent End-to-End Spatiotemporal Feature Learning Framework for Change Detection in Unseen Videos. *IEEE Transactions on Image Processing* 30 (2020), 546–558. https://doi.org/10.1109/TIP.2020.3037472

[82] Larry Matthies, Yalin Xiong, R Hogg, David Zhu, A Rankin, Brett Kennedy, Martial Hebert, R Maclachlan, Chi Won, Tom Frost, et al. 2002. A portable, autonomous, urban reconnaissance robot. *Robotics and Autonomous Systems* 40, 2-3 (2002), 163–172.

[83] Joseph A. Maxwell. 2010. Using Numbers in Qualitative Research. *Qualitative Inquiry* 16, 6 (2010), 475–482. https://doi.org/10.1177/1077800410364740 arXiv:https://doi.org/10.1177/1077800410364740

[84] 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 Transactions on Accessible Computing (TACCESS)* 13, 2 (2020), 1–47.

[85] Bob-Antoine J Menelas, Lorenzo Picinali, Patrick Bourdot, and Brian FG Katz. 2014. Non-visual identification, localization, and selection of entities of interest in a 3D environment. *Journal on Multimodal User Interfaces* 8, 3 (2014), 243–256.

[86] Krista Merry and Pete Bettinger. 2019. Smartphone GPS accuracy study in an urban environment. *PLOS ONE* 14, 7 (07 2019), 1–19. https://doi.org/10.1371/journal.pone.0219890

[87] Vidula V Meshram, Kailas Patil, Vishal A Meshram, and Felix Che Shu. 2019. An astute assistive device for mobility and object recognition for visually impaired people. *IEEE Transactions on Human-Machine Systems* 49, 5 (2019), 449–460.

[88] Microsoft. 2018. Microsoft Soundscape. https://www.microsoft.com/en-us/research/product/soundscape/. [Accessed 09-11-2025].

[89] Microsoft. 2024. Seeing AI - Talking Camera for the Blind. https://www.seeingai.com/. [Accessed 29-08-2025].

[90] Alexandra Millonig and Katja Schechtner. 2007. Developing landmark-based pedestrian-navigation systems. *IEEE Transactions on intelligent transportation systems* 8, 1 (2007), 43–49.

[91] MIPsoft. 2016. BlindSquare. https://www.blindsquare.com/. [Accessed 09-11-2025].

[92] Marko Modsching, Ronny Kramer, and Klaus ten Hagen. 2006. Field trial on GPS Accuracy in a medium size city: The influence of built-up. In *3rd workshop on positioning, navigation and communication*, Vol. 2006. 1–10.

[93] Cecily Morrison, Martin Grayson, Rita Faia Marques, Daniela Massiceti, Camilla Longden, Linda Wen, and Edward Cutrell. 2023. Understanding personalized accessibility through teachable ai: designing and evaluating find my things for people who are blind or low vision. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.

[94] Mahjabin Nahar, Eun-Ju Lee, Jin Won Park, and Dongwon Lee. 2025. Catch Me if You Search: When Contextual Web Search Results Affect the Detection of Hallucinations. *arXiv preprint arXiv:2504.01153* (2025).

[95] 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 preprint arXiv:2404.03745* (2024).

[96] Khanh Hieu Ngo, Quang Khoi Tran, Chan Huy Quan, Nguyen Kim Ngan Luu, Minh Chien Nguyen, Mattis Dautremont, and Anouk Badie. 2023. Photogrammetry-Based Building Scan Using Ultra-Lightweight Drone for Urban Applications. In *The International Conference on Sustainable Civil Engineering and Architecture*. Springer, 1700–1709.

[97] Niantic. 2022. Lightship VPS: Building Our 3D Map From Crowdsourced Scans. https://nianticlabs.com/news/vps-part-2?hl=en.

[98] Niantic. 2024. Building a Large Geospatial Model to Achieve Spatial Intelligence — nianticlabs.com. https://nianticlabs.com/news/largegeospatialmodel?hl=en. [Accessed 14-11-2025].

[99] Niantic. 2025. Contribute to Wayspots across Niantic games. https://wayfarer.nianticlabs.com/new/.

[100] Niantic. 2025. How to Create a Device Map. https://lightship.dev/docs/ardk/how-to/shared_ar/creating_device_maps/ Accessed: 2025-09-05.

[101] Niantic. 2025. How to Create a VPS-Activated Location. Niantic Spatial Platform Developer Documentation. https://lightship.dev/docs/ardk/3.0/how-to/vps/create_vps_activated_location/.

[102] Niantic. 2025. How to Exclude Semantic Channels with Mesh Filtering. https://lightship.dev/docs/ardk/how-to/ar/semantic_mesh_filtering/. Retrieved August 29, 2025.

[103] Niantic. 2025. Niantic SDK for Unity (Lightship ARDK) documentation. https://lightship.dev/docs/ardk/. Retrieved August 28, 2025.

[104] Niantic. 2025. *Visual Positioning System (VPS)*. Niantic. https://www.nianticspatial.com/en/products/visual-positioning-system

[105] Nels Numan, Gabriel Brostow, Suhyun Park, Simon Julier, Anthony Steed, and Jessica Van Brummelen. 2025. CoCreatAR: Enhancing Authoring of Outdoor Augmented Reality Experiences Through Asymmetric Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1232, 22 pages. https://doi.org/10.1145/3706598.3714274

[106] Nels Numan, Jessica Van Brummelen, Ziwen Lu, and Anthony Steed. 2025. AdjustAR: AI-Driven In-Situ Adjustment of Site-Specific Augmented Reality Content. *arXiv preprint arXiv:2508.06826* (2025).

[107] Timo Nummenmaa, Samuli Laato, Philip Chambers, Tuomas Yrttimaa, Mikko Vastaranta, Oğuz Turan Buruk, and Juho Hamari. 2025. Employing Gamified Crowdsourced Close-Range Sensing in the Pursuit of a Digital Twin of the Earth. *International Journal of Human–Computer Interaction* 41, 8 (2025), 4668–4684. https://doi.org/10.1080/10447318.2024.2352922 arXiv:https://doi.org/10.1080/10447318.2024.2352922

[108] OpenAI. 2025. GPT-4o model documentation. https://platform.openai.com/docs/models/gpt-4o. Accessed August 28, 2025.

[109] OpenAI. 2025. GPT-4o-Mini model documentation. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed August 28, 2025.

[110] OpenAI. 2025. A practical guide to building agents. https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf. [Accessed 29-08-2025].

[111] OpenStreetMap contributors. 2017. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

[112] Charlie Parker, Sam Scott, and Alistair Geddes. 2019. Snowball sampling. *SAGE research methods foundations* (2019).

[113] Ricardo E. Gonzalez Penuela, Jazmin Collins, Cynthia L. Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3642211

[114] Per Persson, Fredrik Espinoza, Petra Fagerberg, Anna Sandin, and Rickard Cöster. 2003. GeoNotes: A location-based information system for public spaces. In *Designing information spaces: the social navigation approach*. Springer, 151–173.

[115] Helen Petrie. 2023. *Crowdsourcing descriptions of visual works of art for blind and partially sighted people*. Ph. D. Dissertation. York.

[116] Niantic Spatial Platform. 2025. Visual Positioning System (VPS). https://lightship.dev/docs/ardk/features/lightship_vps/. [Accessed 09-11-2025].

[117] R Prabha, M Razmah, G Saritha, RM Asha, R Gayathiri, et al. 2022. Vivoice-Reading Assistant for the Blind using OCR and TTS. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 01–07.

[118] Sanjana Prasain. 2011. StopFinder: improving the experience of blind public transit riders with crowdsourcing. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. 323–324.

[119] Yue Qiu, Shintaro Yamamoto, Ryosuke Yamada, Ryota Suzuki, H. Kataoka, K. Iwata, and Y. Satoh. 2023. 3D Change Localization and Captioning from Dynamic Scans of Indoor Scenes. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2023), 1176–1185. https://doi.org/10.1109/WACV56688.2023.00123

[120] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Whisper. https://github.com/openai/whisper. OpenAI, Automatic Speech Recognition System.

[121] Adil Rahman, Md Aashikur Rahman Azim, and Seongkook Heo. 2023. Take my hand: Automated hand-based spatial guidance for the visually impaired. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[122] Aryan Mukesh Rajpurohit, Puneet Kumar, Dalwinder Singh, and Rishu Raj Kumar. 2023. A Review on Visual Positioning System. In *Proceedings of the KILBY 100 7th International Conference on Computing Sciences*.

[123] Sanjeev U Rao, Swaroop Ranganath, TS Ashwin, Guddeti Ram Mohana Reddy, et al. 2021. A Google glass based real-time scene analysis for the visually impaired. *IEEe Access* 9 (2021), 166351–166369.

[124] Matthew T Rice, R Daniel Jacobson, Douglas R Caldwell, Scott D McDermott, Fabiana I Paez, Ahmad O Aburizaiza, Kevin M Curtin, Anthony Stefanidis, and Han Qin. 2013. Crowdsourcing techniques for augmenting traditional accessibility maps with transitory obstacle information. *Cartography and Geographic Information Science* 40, 3 (2013), 210–219.

[125] John TE Richardson. 2018. The use of Latin-square designs in educational and psychological research. *Educational Research Review* 24 (2018), 84–97.

[126] R. Russell, A. Guerry, P. Balvanera, Rachelle K. Gould, X. Basurto, K. Chan, Sarah C. Klain, J. Levine, and Jordan Tam. 2013. Humans and Nature: How Knowing and Experiencing Nature Affect Well-Being. *Annual Review of Environment and Resources* 38 (2013), 473–502. https://doi.org/10.1146/ANNUREV-ENVIRON-012312-110838

[127] 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*. 222–235.

[128] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. 2019. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[129] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human*

Computation and Crowdsourcing, Vol. 5. 147–156.
[130] 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 Transactions on Accessible Computing (TACCESS)* 12, 3 (2019), 1–30.
[131] Filipo Sharevski and Aziz Zeidieh. 2023. Designing and conducting usability research on social media misinformation with low vision or blind users. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop*. 75–81.
[132] Filipo Sharevski and Aziz N Zeidieh. 2023. "I Just Didn't Notice It:" Experiences with Misinformation Warnings on Social Media amongst Users Who Are Low Vision or Blind. In *Proceedings of the 2023 New Security Paradigms Workshop*. 17–33.
[133] Huiying Shen, Owen Edwards, Joshua Miele, and James M Coughlan. 2013. CamIO: a 3D computer vision system enabling audio/haptic interaction with physical objects by blind users. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–2.
[134] Lee Stearns, Leah Findlater, and Jon E Froehlich. 2018. Design of an augmented reality magnification aid for low vision users. In *Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility*. 28–39.
[135] Kai Sun, Yingjie Hu, Yue Ma, Ryan Zhenqi Zhou, and Yunqiang Zhu. 2023. Conflating point of interest (POI) data: A systematic review of matching methods. *Computers, Environment and Urban Systems* 103 (2023), 101977. https://doi.org/10.1016/j.compenvurbsys.2023.101977
[136] 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*. 61–72.
[137] TapTapSee. n.d.. *TapTapSee: Assistive Technology for the Blind and Visually Impaired*. TapTapSee. https://taptapseeapp.com/
[138] Gareth Terry, Nikki Hayfield, Victoria Clarke, Virginia Braun, et al. 2017. Thematic analysis. *The SAGE handbook of qualitative research in psychology* 2, 17-37 (2017), 25.
[139] Jessica Van Brummelen, Liv Piper Urwin, Oliver James Johnston, Mohamed Sayed, and Gabriel Brostow. 2024. Don't Look Now: Audio/Haptic Guidance for 3D Scanning of Landmarks. 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 415, 20 pages. https://doi.org/10.1145/3613904.3642271
[140] Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. 95–102.
[141] Ru Wang, Zach Potter, Yun Ho, Daniel Killough, Linxiu Zeng, Sanbrita Mondal, and Yuhang Zhao. 2024. GazePrompt: Enhancing Low Vision People's Reading Experience with Gaze-Aware Augmentations. *arXiv preprint arXiv:2402.12772* (2024).
[142] Wenwen Wei, Ping Wei, Jialu Qin, Zhimin Liao, Shuaijie Wang, Xiang Cheng, Meiqin Liu, and Nanning Zheng. 2023. 3d scene graph generation from point clouds. *IEEE Transactions on Multimedia* 26 (2023), 5358–5368.
[143] Hans Jørgen Wiberg. 2015. Be My Eyes. https://www.bemyeyes.com/. Launched in 2015; accessed 2025-06-25.
[144] Walter Wittich, Jonathan Jarry, Elliott Morrice, and Aaron Johnson. 2018. Effectiveness of the Apple iPad as a spot-reading magnifier. *Optometry and Vision Science* 95, 9 (2018), 704–710.
[145] Jianfeng Wu. 2023. VoiceVista. https://apps.apple.com/us/app/voicevista/id6450388413 Accessed: 2025-09-06.
[146] X Help Center. [n. d.]. About X's verified accounts. https://help.x.com/en/managing-your-account/about-x-verified-accounts. Accessed: 2025-09-03.
[147] Demet Yesiltepe, Ruth Conroy Dalton, and Ayse Ozbil Torun. 2021. Landmarks in wayfinding: a review of the existing literature. *Cognitive processing* 22 (2021), 369–410.
[148] Chris Yoon, Ryan Louie, Jeremy Ryan, MinhKhang Vu, Hyegi Bang, William Derksen, and Paul Ruvolo. 2019. Leveraging augmented reality to create apps for people with visual disabilities: A case study in indoor navigation. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 210–221.
[149] Rui Yu, Sooyeon Lee, Jingyi Xie, Syed Masum Billah, and John M Carroll. 2024. Human–AI collaboration for remote sighted assistance: Perspectives from the LLM era. *Future internet* 16, 7 (2024), 254.
[150] Tianhao Yu. 2024. *Automated drone navigation for urban data collection*. Ph. D. Dissertation. University of Illinois at Urbana-Champaign.
[151] Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. 2024. Explaining the Wait: How Justifying Chatbot Response Delays Impact User Trust. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 27, 16 pages. https://doi.org/10.1145/3640794.3665550

A Participant Demographics in the Formative Study

B Participant Demographics in the Evaluation

C Note on the Appendices

Please note that in addition to the above appendices, this paper has a separate Supplementary Materials document.

Table 1: Participant demographics in the formative study.

ID	Day and Group	Age/Gender	Visual Condition	"Can you match clothes/objects according to their color?"	"Can you notice objects around you as you walk?"	If you were paired with a totally blind participant, would you be able to help guide them?
F1	Day 1, Group A	45/F	Low vision	Sometimes	When close by	N
F2	Day 1, Group A	43/M	Low vision	TapTapSee tells me colors of things	Yes with good contrast	Sometimes
F3	Day 1, Group A	20/F	Low vision	N	If it's far then no	Y
F4	Day 1, Group A	33/F	Low vision, macular dystrophy	Y	If it's near yes but if it's far it harder to see	Y
F5	Day 1, Group B	58/F	Totally blind	N	N	N
F6	Day 1, Group B	25/F	Low vision	Y	Y	Y
F7	Day 1, Group B	51/F	Low vision	Sometimes	Y	Y
F8	Day 1, Group B	56/F	Low vision	N	Sometimes	Sometimes
F9	Day 1, Group C	46/F	Totally blind	With help	Using cane	N
F10	Day 1, Group C	61/F	Low vision	Sometimes	Depends on distance	Y
F11	Day 1, Group C	45/M	Low vision	N	N	Not answered
F12	Day 1, Group C	45/M	Low vision	Depends on color	Depends on light	Y
F13	Day 2, Group D	44/F	Low vision	Y	N	Not answered
F14	Day 2, Group D	60/F	Low vision	Sometimes or most times, but it has to be at very close proximity and it's dependent on the level of light	Sometimes. It depends how close it is and only in daylight.	Sometimes, it depends on the light and if the ground is flat.
F15	Day 2, Group D	76/M	Low vision	N	N	Y
F16	Day 2, Group D	24/F	Low vision, peripheral vision loss	Depends	Depends	N
F17	Day 2, Group E	48/F	Low vision, color blind, peripheral vision loss	N	Not answered	N
F18	Day 2, Group E	62/M	Low vision, no sight in left eye	Y	Y	Y
F19	Day 2, Group E	66/M	Low vision	Y	Y	Y
F20	Day 2, Group E	58/F	Low vision	No. I can see light and shapes	Yes. I can see objects in blurry vision	N
F21	Day 2, Group F	36/F	Low vision	Y	Sometimes but not in the dark	Y
F22	Day 2, Group F	40/M	Low vision	Y	Y	Y
F23	Day 2, Group F	34/M	Blind with some light perception	N	N	Not answered
F24	Day 2, Group F	65/M	Low vision	Y	Blurry vision	Y

Table 2: Participant demographics in the evaluation.

ID	Formative ID	Age/Gender	Visual Condition	How Much They Rely on Vision in Daily Activities	Prior AR Experiences	How do you typically navigate the last 10-50 meters when getting to a destination?
Pilot1	F7	51/F	Low vision, central vision loss	Always	N	Not asked
Pilot2	F19	66/M	Low vision	Always	Y	Not asked
Pilot3	F4	58/F	Blind	Not at all	Y	Not asked
Pilot4	F23	35/M	Low vision	Sometimes	Y	Not asked
P1	F3	20/F	Low vision, central vision loss	Always	Y	A sighted guide (e.g., friend, family member), asking for verbal directions from people nearby, a general navigation app (e.g., Google Maps, Apple Maps)
P2	F15	78/M	Totally blind	Not at all	N	A white cane, a sighted guide
P3	F16	25/F	Low vision	Sometimes	N	Asking for verbal directions from people nearby, a real-time remote assistance app (e.g. Be My Eyes Volunteer, Aira Agent)
P4	F4	33/F	Low vision, sensitive to light, central vision loss	Sometimes	Y	A sighted guide, asking for verbal directions from people nearby, a general navigation app
P5	-	35/F	Blind with some light perception	Rarely	N	A white cane, a sighted guide, asking for verbal directions from people nearby, a real-time remote assistance app
P6	F8	56/F	Low vision	Sometimes	N	A white cane, a sighted guide, asking for verbal directions from people nearby, a real-time remote assistance app
P7	-	55/M	Low vision, sensitive to light, weak depth perception, central vision loss, scotoma	Always	Y	A white cane, a sighted guide, asking for verbal directions from people nearby, a real-time remote assistance app
P8	-	46/M	Totally blind	Not at all	N	A sighted guide, Asking for verbal directions from people nearby, a general navigation app, a BLV or accessibility app I can use independently (e.g., Seeing AI, VoiceVista, Be My Eyes without remote assistance)
P9	-	80/M	Totally blind	Not at all	N	A white cane, asking for verbal directions from people nearby, a real-time remote assistance app
P10	F2	44/M	Low vision, peripheral vision loss, scotoma, occasionally sensitive to light	Always during daylight, cannot see at night	Y	A white cane, a sighted guide, Asking for verbal directions from people nearby, a real-time remote assistance app, a BLV or accessibility app I can use independently
P11	-	66/F	Low vision, weak depth perception, peripheral vision loss	Sometimes	Y	A white cane, a sighted guide, Asking for verbal directions from people nearby, a real-time remote assistance app
P12	F1	46/F	Low vision, central vision loss	Always	Y	A white cane, A sighted guide, asking for verbal directions from people nearby, a general navigation app, a real-time remote assistance app, a BLV or accessibility app I can use independently
P13	-	42/F	Blind with some light perception, color blind, no depth perception	Not at all	N	A white cane, asking for verbal directions from people nearby
P14	-	67/M	Low vision, sensitive to light, color blind, weak depth perception, peripheral vision loss	Always	Y	A white cane, a sighted guide, asking for verbal directions from people nearby
P15	-	50/F	Low vision	Always	Y	A white cane, a sighted guide, asking for verbal directions from people nearby, a general navigation app. If its really important I might do the journey on an earlier day so I don't get lost on the important day. Also I often arrange someone to meet me at a place I know I can find like a tube station.
P16	F21	37/F	Low vision, color blind, peripheral vision loss	Always	N	A sighted guide, asking for verbal directions from people nearby, a general navigation app
P17	-	49/M	Blind with some light perception	Just to identify day and night	N	A white cane, asking for verbal directions from people nearby
P18	-	52/F	Blind with some light perception	Rarely	N	A white cane, a sighted guide