International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

ObjectFinder: An Open-Vocabulary Assistive System for Interactive Object Search by People Who Are Blind

Ruiping Liu, Jiaming Zhang, Angela Schön , Karin Müller, Junwei Zheng, Kailun Yang, Anhong Guo, Kathrin Gerling & Rainer Stiefelhagen

To cite this article: Ruiping Liu, Jiaming Zhang, Angela Schön , Karin Müller, Junwei Zheng, Kailun Yang, Anhong Guo, Kathrin Gerling & Rainer Stiefelhagen (06 Jul 2026): ObjectFinder: An Open-Vocabulary Assistive System for Interactive Object Search by People Who Are Blind, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2695932

To link to this article: https://doi.org/10.1080/10447318.2026.2695932

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2695932

ObjectFinder: An Open-Vocabulary Assistive System for Interactive Object Search by People Who Are Blind

Ruiping Liu[a], Jiaming Zhang[b], Angela Schön[a], Karin Müller[a], Junwei Zheng[a], Kailun Yang[b], Anhong Guo[c], Kathrin Gerling[a] and Rainer Stiefelhagen[a]

[a]Karlsruhe Institute of Technology, Karlsruhe, Germany; [b]Hunan University, Changsha, China; [c]University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Searching for objects in unfamiliar scenarios is a fundamental challenge for people with blindness, requiring target specification, detection, and intent-specific details, such as navigating toward the object and understanding its surroundings. However, existing description- and detection-based assistive technologies fail to address this multifaceted nature of interactive object search. We present ObjectFinder, a wearable prototype with smart glasses (RGB-D sensing), a bone conduction headset, and an edge computing unit for interactive object search by people with blindness. Users query target objects with flexible wording; once detected, ObjectFinder provides real-time egocentric localization information and lets usersinitiate different branches to gather intent-specific details. ObjectFinder integrates open-vocabulary object detection with a multimodal large language model into an intent-driven pipeline, developed iteratively with a blind co-designer. In a study with eight participants with blindness, ObjectFinder's flexible target specification, egocentric and allocentric cues, and targeted context around the target were valued, with lower NASA-TLX task load than the baseline.

KEYWORDS

Assistive technology; users with blindness; object search; human–AI interaction; wearable systems

1. Introduction

People with blindness often face challenges when searching for objects in unfamiliar environments (Müller et al., 2022; Zhang et al., 2021), such as independently searching for far away objects (e.g., an information desk in a spacious lobby (Constantinescu, Müller, et al., 2022)), identifying and understanding objects with insufficient tactile cues (e.g., finding the shampoo bottle among similar containers in an unfamiliar hotel bathroom (Brady et al., 2013)), and searching for items on large surfaces that are inconvenient to explore by touch (e.g., finding specific tools on a cluttered tabletop (Herskovitz et al., 2023)).

To search for an object, users need to ask for the target object, and then select a candidate. In order to determine if it is the desired one, both egocentric (e.g., "*11 o'clock, 2.4 meters away*") and allocentric (e.g., "*next to the desk*") information are necessary for a user with blindness to understand the object in its surroundings (Martolini et al., 2021; Reginald, 1999). Upon locating the target object, the user's intent may vary. As shown in examples of Figure 1, once the coffee table is detected, the user may prefer a description of the items on the table over immediate physical interaction. Conversely, if the target is a fan, the user might wish to navigate toward it to turn it on. Throughout the search, there may also be an interest to explore the surroundings for better navigation and potentially discover other targets for subsequent searches (Jain, Teng, et al., 2023).

Such an interactive object search task is multifaceted. While assistive technology, broadly defined as products and systems that maintain or improve an individual's functioning (World Health Organization, 2016), has been increasingly applied to support people who are blind in understanding

CONTACT Ruiping Liu ruiping.liu@kit.edu Karlsruhe Institute of Technology, Karlsruhe, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2026.2695932.

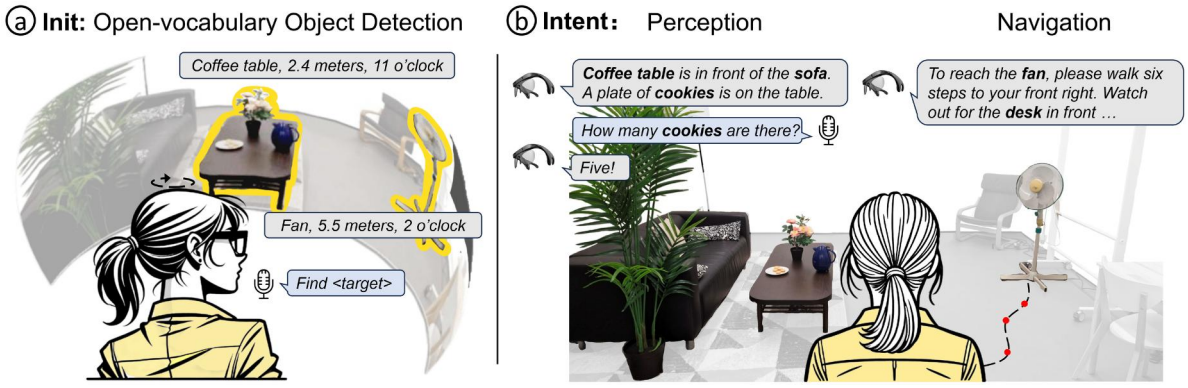

Figure 1. ObjectFinder system for open-vocabulary interactive object search. It seamlessly integrates open-vocabulary models, i.e., an open-vocabulary object detector (e.g., YOLO-World) and a multimodal large language model (e.g., GPT-4). (a) a user specifies a target with flexible wording on smart glasses. Once it is found, the user is informed with egocentric localization information in real-time. (b) Upon detecting the target object, the user may have various intentions toward it, such as perceiving what is on the coffee table or navigating toward a fan to turn it on. During the interaction, the user may discover other objects of interest for subsequent searches, e.g., cookies on the coffee table.

their surroundings, no current system provides a unified solution that handles all associated subtasks. We categorize the existing assistive technologies for people with blindness into *description-based* and *detection-based* systems. Description-based systems can provide detailed descriptions of photos (BeMyAI, 2023), brief captions (Lee, Herskovitz, et al., 2022; Liu et al., 2023; Microsoft Corporation, 2024), or descriptions with dynamic granularity and latency (Chang et al., 2024). However, difficulty in aiming the phone camera limits these systems' ability to localize a specific object in unfamiliar environments (Chang et al., 2024; Xie et al., 2024). **(Challenge 1, C1)**. Detection-based systems (Google, 2024; V7, 2024; Zheng et al., 2024), on the other hand, either allow only the search for a limited number of pre-defined objects (Ahmetovic et al., 2020; Constantinescu et al., 2020; Google, 2024; Kacorri et al., 2017; Schauerte et al., 2012; V7, 2024; Wen et al., 2024; Yi et al., 2013) or provide filtered information (Herskovitz et al., 2024), limiting understanding of unfamiliar scenes. Therefore, when using current detection-based systems in an unfamiliar environment, users with blindness may not know what is in a room comprehensively and miss items that could be of interest **(Challenge 2, C2)**.

The challenges also exist in Remote Sighted Assistance (RSA). The procedure by which the remote agents (Aira Company, 2024; BeMyEyes, 2024; TapTapSee, n.d) help to identify objects and describe surroundings involves capturing images from the video feed and zooming in to obtain the necessary visual information (Lee et al., 2020). In this context, it is time-consuming for remote agents to adjust the video frame, and they find it challenging to continuously orient the users (Xie, Reddie, et al., 2022). Moreover, recognizing landmarks presents significant difficulties for the agents (Kamikubo et al., 2020; Lee, Reddie, et al., 2022; Xie, Yu, et al., 2022). Thus, in this work, we aim to investigate:

How to integrate the advantages of description- and detection-based assistive systems to support interactive object search by people with blindness?

To this end, we designed ObjectFinder, which seamlessly combines open-vocabulary models, an Open-Vocabulary Object Detector (OVOD) (Cheng et al., 2024), and a Multimodal Large Language Model (MLLM) (Achiam et al., 2023), to facilitate an interactive process that ranges from object detection to description for object search. Users can input any target using voice commands for object detection, and then scan the scene. Once a candidate is detected, the system will notify the user to stand still to orient to the target, and it will provide real-time egocentric information (distance and direction). Following this, users can acquire targeted scene context tailored to their post-search intent, based on the keyframe captured at the time of detection. This process helps users recognize candidate options and unexpected targets, which can then be explored in further detail during subsequent iterations. The subtask definition and pipeline refinement were conducted with a co-designer who is blind across two months and four iterations.

To evaluate the effectiveness of ObjectFinder in supporting interactive object search, we conducted a user study with eight participants with blindness. The results show that the targeted scene context enabled by the seamless integration of detection and description enhanced not only the search process itself but also the post-search intentions. It provided essential information for object search, including egocentric localization (distance and direction), and allocentric relationships among objects. Additionally, route planning was a valuable feature of ObjectFinder for searching objects. Participants valued the ability to specify targets flexibly. Although ObjectFinder provides feedback based on users' intents, individual variations in procedures and information preferences, also influenced by the scope and familiarity of the scenario, underscore the need for customization and personalization, as discussed later. Despite the rich information provided by the MLLM, the 79.8% direction accuracy in route planning and 82.5% spatial accuracy in scene description observed in the original system highlight the need for continued improvements in spatial reasoning capabilities. The system's code will be made publicly available to contribute to the open-source development of assistive technologies (Buehler et al., 2015). The contribution of this study can be summarized as follows:

- To the best of our knowledge, ObjectFinder is the first system to seamlessly integrate open-vocabulary object detection with a multimodal large language model, enabling a flexible, intent-driven object-search experience for people with blindness in unfamiliar environments.
- We iteratively developed the pipeline with a co-designer with blindness. In a study with eight participants who are blind, users reported a new object-search experience in which they could flexibly specify targets, stay oriented with egocentric and allocentric cues, and receive targeted scene context supporting both search and post-search intents.
- Beyond the technical implementation, this work contributes design insights and empirical findings that can inform the broader development of AI-powered assistive technologies balancing flexibility, reliability, and user agency.

2. Related work

In this section, we introduce the task of object search for individuals with blindness and provide an overview of existing description-based and detection-based systems designed specifically for them, which can partially address the task. To design an assistive system for object search in unfamiliar environments, we draw on procedures from embodied AI, which typically model human search behavior. This forms the background knowledge for our study.

2.1. Object search in unfamiliar scenarios

Object search is a multifaceted task that involves object detection, exploration, navigation, and more. In addition to small items that individuals with blindness frequently search for in daily life, such as smartphones, keys, and wallets (netz-barrierefrei.de, n.d), they often look for large, salient objects as landmarks to improve orientation in unfamiliar environments (ATMAPS Project, n.d; Yang et al., 2010). When searching for objects in unfamiliar environments, people with blindness typically seek an initial overview of the space, followed by specific details as required (Chang et al., 2024; Shneiderman, 1996). If the target object has been found, people with blindness may have various intentions regarding it. For example, they might navigate to the object to interact with it (Herskovitz et al., 2023) (e.g., find a free chair and sit on it), identify a specific object (Brady et al., 2013; Hong & Kacorri, 2024a) (e.g., check whether a bottle is shampoo), or perceive the surroundings of the object (e.g., the tabletop (Herskovitz et al., 2023)), which may be too far or inconvenient to touch (Gonzalez Penuela et al., 2024). A participant in (Herskovitz et al., 2023) described locating an empty chair in a classroom and envisioned an object search system that detects the chair via smart glasses and provides walking directions, avoiding the need to wave a phone in a crowded space.

Some technologies have been proposed to partially address the challenges of object search (Table 1). Vizwiz-LocateIt (Bigham et al., 2010) lets users photograph target objects, ask questions to a remote worker on Mechanical Turk, and navigate via sonification. Tools such as AIRA (Aira Company, 2024),

Vizwiz (Gurari et al., 2018), and BeMyEyes (2024) use crowdsourcing to connect users with blindness with sighted assistants for real-time remote support, including object search. However, repeatedly asking users to move their phone to adjust the camera view is time-consuming (Lee et al., 2020). WanderGuide (Kuribayashi et al., 2025) has subfunctions for object search implemented on a suitcase, but is designed primarily for exploration without specific consideration of the object search procedure. Gamage et al. (2023) suggest that the conversational interface on wearable devices is suitable for the complex task of providing environmental information. We categorize existing AI-based assistive technology related to object search into description-based and detection-based systems.

2.2. Description-based systems for people with blindness

Description-based systems typically capture a still image and generate description for it. Seeing AI (Microsoft Corporation, 2024), ImageExplorer (Lee et al., 2022a), and OpenSU (Liu et al., 2023) provide brief captions for scenes captured by a mobile phone and support tactile exploration on touchscreens. BeMyAI (2023) powered by GPT-4, offers richer scene descriptions and interactive question answering; however, prior studies (Xie et al., 2024, 2025) show that it struggles with intent recognition and often requires multiple images to convey accurate information. NaviGPT (Zhang et al., 2025) focuses on navigation by describing the road ahead, while LifeInsight (Mathis & Schöning, 2025) extends GPT-4–based question answering to wearable settings. Other work targets specific object attributes, such as material (Zheng et al., 2024), transparency (Zhang et al., 2021), or hazards (Yang, Bergasa, Romera, Cheng, et al., 2018; Yang, Bergasa, Romera, Huang, et al. 2018). However, description-based applications require users with blindness to aim and maintain the camera toward the target object, which is difficult in unfamiliar environments (Gurari et al., 2018; Yang, He, et al., 2024). In our user study, participants with blindness reported that photo descriptions did not help them infer where objects were located in real space (e.g., relative direction or distance), consistent with prior findings (Kacorri et al., 2017).

WorldScribe (Chang et al., 2024) provides dynamic scene descriptions by concurrently integrating YOLO-World (Cheng et al., 2024) and GPT-4 (Achiam et al., 2023) to describe the current camera view, rather than to support targeted object search. YOLO-World offers fast word-level cues, while GPT-4 generates sentence-level descriptions; user intent (e.g., *"find the laptop"*) is only used to rank mentioned content for open-ended exploration, not to localize a specific target. Although WorldScribe can partially support object search, prior findings report that users may struggle to aim and maintain the camera view and that essential objects may be omitted. Guided by the task–artifact cycle theory (Carroll & Rosson, 1992), we argue that task-specific designs provide distinct value. In contrast, ObjectFinder is explicitly goal-oriented: YOLO-World first localizes the target object, after which GPT-4 supplies targeted scene context aligned with the user's post-search intent, e.g., route planning for navigation or focused scene description for perception.

2.3. Detection-based systems for people with blindness

Detection-based systems are designed to provide real-time outputs of identified objects or features of interest. Lookout (Google, 2024), AIPoly (V7, 2024), and Supersense (2024) exemplify this capability by identifying the nearest object within the phone's field of view. Various studies have developed wearable systems (Islam et al., 2023; Liu, 2023; Sugashini & Balakrishnan, 2024) with similar functionalities, offering real-time object information through multiple interaction modes. Research has explored the detection of personal objects using methods such as SIFT (Constantinescu et al., 2020; Schauerte et al., 2012; Yi et al., 2013) and advanced deep learning networks (Ahmetovic et al., 2020; Kacorri et al., 2017; Wen et al., 2024). ProgramAlly (Herskovitz et al., 2024) allows users to customize filters to detect specific object features, suggesting that fixed, predefined object-category lists are less suitable for open-ended exploration. Recent work has further incorporated open-vocabulary detection into wearable assistive systems; for instance, Feng et al. (2026) integrate YOLO-World into a pipeline for construction site hazard detection for people who are blind. However, their system targets a fixed hazard category rather than user-specified objects during live use. Navigation assistive systems utilize detection-based

Table 1. Comparison of different systems that can be used for object search.

System	Purpose	Enabling source	Interaction	Device
RSA (e.g., Aira Company, 2024, BeMyEyes, 2024)	Multi-Purpose	Human	Dialogue	Smartphone
ProgramAlly (Herskovitz et al., 2024)	Object Search	AI	Filter Customization	Smartphone
WorldScribe (Chang et al., 2024)	Exploration	AI	Intent Customization	Smartphone
WanderGuide (Kuribayashi et al., 2025)	Exploration	AI	Dialogue, Button-Driven Option Selection	Robot
BeMyAI (2023)	Description	AI	Dialogue	Smartphone
Lookout (Google, 2024)	Exploration	AI	–	Smartphone
Find My Things (Wen et al., 2024)	Object Search	AI	Teachable Object Recognition	Smartphone
LifeInsight (Mathis & Schöning, 2025)	Question and Answering	AI	Dialogue	Wearable Device
ObjectFinder	Object Search & Exploration	AI	Dialogue, Button-Driven Option Selection	Wearable Device

methods for obstacle avoidance (Bala et al., 2023; Liu et al., 2021; Ou et al., 2022), risk assessment (Wang et al., 2024), object finding (Duh et al., 2020; Hu et al., 2022; Li et al., 2023), shopping (Boldu et al., 2020) and passable path planning (Hong et al., 2024; Jain et al., 2023a; Sorougi & McCann, 2023; Zou et al., 2023). These systems autonomously select information, which may limit user agency in actively specifying targets. Constantinescu et al. (2020) propose a system that allows users to choose from a limited set of objects to receive audible feedback in their vicinity. Detection-based systems typically report only information within a predefined range, such as objects from a fixed list, personal objects, or feature-based filters, which prevents broader scene understanding during object search and leaves them vulnerable to failures under challenging real-world conditions such as poor lighting, occlusion, or cluttered backgrounds.

2.4. Reference procedure for object search

Since no existing AI system fully addresses the challenges of object search for users with blindness, we examine the object-search procedures of embodied AI to guide the design of an assistive object search system. Object search is widely recognized as a challenging task that integrates both perceptual and cognitive processes (Sun et al., 2025). Typically, embodied agents (Aydemir et al., 2013; Singh Chaplot et al., 2020; Yokoyama et al., 2024) first receive an object query from the user, analyze their surroundings, hypothesize the potential location of the target object, and then plan a navigation path accordingly. Recent workflows leveraging LLMs, such as UniGoal (Yin et al., 2025) and SG-Nav (Yin et al., 2024), allow robots to continuously explore their environment and match discovered objects with the intended target. CogNav (Cao et al., 2024) models the cognitive process of object search, in which an agent repeatedly explores the environment in a broad, contextual manner to build a cognitive map. Once a potential target is observed, the agent verifies the candidate using information from the surrounding context before confirming it as the target object. Taioli et al. (2024) equipped the object search agent with a self-questioning module and an interaction trigger, enabling it to refine its detections through dialogue about the target object. In our work, we explore the multifaceted subtasks involved in object search and integrate them into a unified pipeline specifically designed for users with blindness, taking advantage of both description- and detection-based systems.

3. ObjectFinder

ObjectFinder is a wearable prototype designed for interactive object search. Users with blindness can specify their target using flexible wording. Once the target is detected, they receive real-time egocentric localization information. They can further obtain detailed feedback based on their intentions toward the targets. We co-designed ObjectFinder with a person with blindness P0 (see Table 2). Firstly, the co-designer proposed an envisioned use scenario for *searching for a socket*, which we used to build an initial prototype. We then conducted four refinement iterations over two months, each a 30-minute

Table 2. Demographics of participants.

User ID	Gender	Age range	Vision level, onset	Experience of apps
P0	Male	30–39	Light perception, since about 2004	BeMyAI, Seeing AI
P1	Female	20–29	Light perception, since about 2022	BeMyAI, Seeing AI
P2	Male	50–59	Fully blind, since birth	BeMyAI, Seeing AI
P3	Male	20–29	Fully blind, since birth	BeMyAI
P4	Male	20–29	Fully blind, since 2010	Lookout
P5	Female	30–39	Light perception, since birth	Seeing AI, Envision
P6	Female	50–59	Fully blind, since about 1989	Seeing AI
P7	Male	70–79	Fully blind, since birth	None
P8	Male	30–39	Light perception, since 2015	Seeing AI

P0 is the co-designer who helped to adapt the system to the needs of the target group. P1–P8 were participants of the user study.

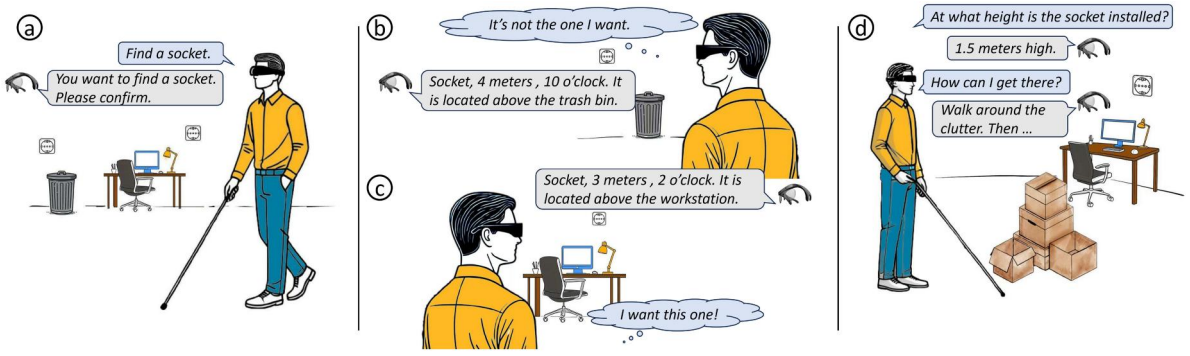

Figure 2. Initial concept of an envisioned assistive object search system and its usage scenario. Martin walks into an unfamiliar office and uses an object-search system to search for a socket to charge his smartphone. (a) Martin first specifies the target to the system, which then repeats it for confirmation. (b) While scanning, candidates are detected. The socket "*4 meters away at his 10 o'clock next to the trash bin*" is not what he wants. (c) However, another socket "*3 meters away at his 2 o'clock next to the workstation*" is the desired one, as he plans to study there. (d) After confirming the target, Martin may ask for more details. In large rooms, the system should navigate him to the socket.

session where P0 freely explored the living room (Figure 8a) and office (Figure 8b) settings with the iteratively refined ObjectFinder.

3.1. Envisioned scenario

In an initial step, we began by defining the use case according to the principles outlined in Cockburn (2000). To achieve this, we conducted a workshop involving the co-designer with blindness, a developer, and two experts in accessibility and usability, one of whom is blind.

The co-designer suggested a use case for object search: *searching for a socket in an unfamiliar office*. We refined the use case regarding the interaction sequence between the co-designer and the wearable object search system as the envisioned scenario (Carroll, 1995), serves as the basis for our system design, as illustrated in Figure 2 and depicted as follows:

Martin enters an unfamiliar office, his phone battery depleted. In need of power, Martin activates an object search system and commands it to *"Find a socket."* The system acknowledges the command, and assures Martin that it has understood the request. After Martin confirms, the system signals with a sound, indicating readiness to begin searching (Figure 2a).

As Martin scans the room through the system, he prefers not to be bombarded with information about every detected object; instead, he wants the system to announce only when it detects a *socket*. Upon identifying a *socket*, the system provides feedback on its egocentric location, including distance and direction, as well as its allocentric relationship with points of interest.

Using this information, Martin evaluates the suitability of the socket's location. In Figure 2b, the first socket detected, located near a trash bin about 4 meters away at a 10 o'clock direction, is deemed inconvenient because Martin intends to work at a workstation. Therefore, he continues his search for a more suitably placed socket.

Eventually, a socket near a workstation, just 3 meters away at the 2 o'clock direction, catches Martin's interest, as in Figure 2c. He requests more details about this socket, such as how to reach it and its height on the wall. The system advises Martin to navigate around obstacles, guiding him with instructions like, "*Walk around the clutter…* " (Figure 2d).

Using his cane to detect and avoid clutter, Martin reaches the workstation located on his front right and successfully charges his phone using the nearby socket.

Five functions can be inferred from the envisioned scenario: (F1) object detection to detect a socket, (F2) localization to provide its egocentric position, (F3) navigation guidance (route planning) for large spaces, (F4) scene description to convey allocentric context around the socket, and (F5) open-ended querying to obtain additional details (e.g., the socket's height). Together, these functions allow the co-designer to verify whether a detected socket is the desired one and then obtain targeted scene context to support subsequent intentions or movement. These functions are modularized and integrated into a flexible, intent-driven pipeline to support interactive object search.

3.2. Design goals

In general, drawing on related work on object search and the envisioned scenario, we designed ObjectFinder with three primary goals:

G1-Providing flexibility in target queries and information retrieval. According to the envisioned scenario for ObjectFinder and the expected procedure for object search in Herskovitz et al. (2023), users would like to directly query a target object. Generally, conversational interfaces are preferred for conveying environmental information (Gamage et al., 2023) and supporting goal-oriented visual querying (Mathis & Schöning, 2025). ObjectFinder should facilitate seamless conversational interactions using open-vocabulary models to bridge wording gaps.

G2-Supporting various subtasks during search. According to the envisioned scenario, object search is a complex task with sequential subtasks including target specification, detection, and feedback generation, which aligns with the expected procedure (Herskovitz et al., 2023). This structure is also reflected in how remote sighted assistants support object search (Lee et al., 2020; Xie et al., 2024) and in embodied systems that mimic human cognition for object search (Aydemir et al., 2013; Singh Chaplot et al., 2020; Yokoyama et al., 2024). ObjectFinder should simplify this process by organizing these subtasks into a user-friendly pipeline with accessible interaction features.

G3-Adapting to the user intent of the target object. People with blindness may have diverse intents after a target is identified, including navigation, e.g., going to the socket for charging in the envisioned scenario, perception, e.g., understanding a tabletop (Herskovitz et al., 2023), and information querying, e.g., asking about the state of household systems (Turkstra et al., 2025). These intents require targeted scene context (Buehler et al., 2015; Hong & Kacorri, 2024a). Prior work suggests that systems should enable users to efficiently articulate and refine evolving intents (Chang et al., 2024). ObjectFinder should therefore provide efficient interactions that allow users to obtain intent-specific feedback from the system.

3.3. Hardware and interaction features

According to the related work (Gamage et al., 2023; Herskovitz et al., 2023) and the envisioned scenario with the co-designer, a pair of glasses with a camera is assumed to be preferred over holding a smartphone for an object search system. We utilize the following hardware to implement this system. Figure 3 presents the system diagram, which comprises a pair of KRVision smart glasses (KR Vision, n.d) coupled with a waist bag. The smart glasses are equipped with an Intel RealSense R200 RGB-D camera, enabling real-time egocentric acquisition of RGB and depth frames at a resolution of 640×480 at 30 FPS, with a depth horizontal field of view of 70°. In addition, a bone conduction headset is incorporated, enabling auditory output while maintaining the perception of environmental sounds. In the waist bag, an NVIDIA Jetson Nano, a compact and powerful processor, is utilized for efficient data processing, accompanied by a power bank for energy supplementation. The waist bag features two

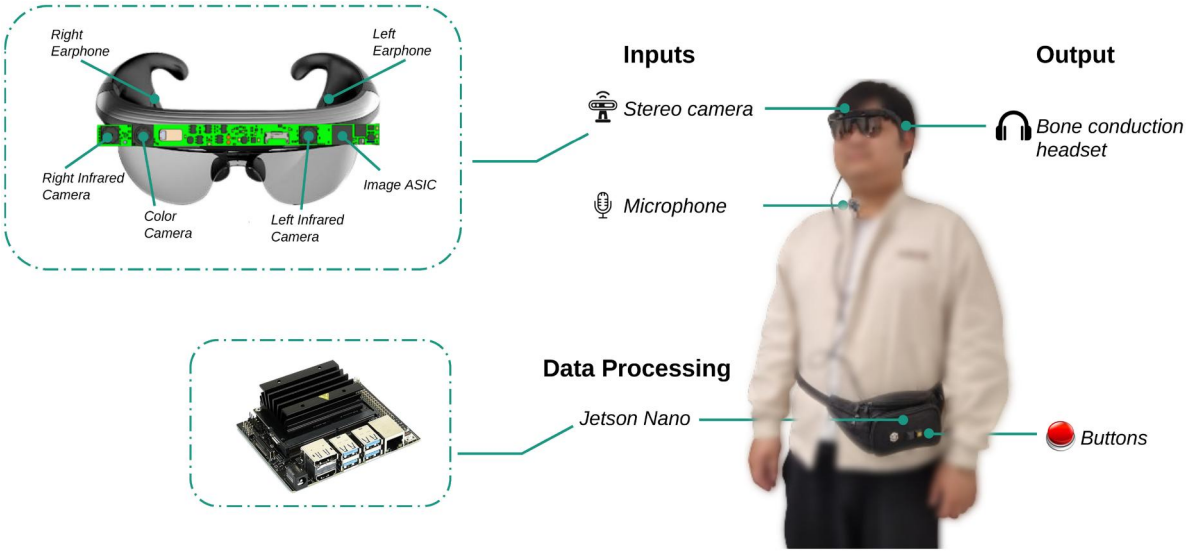

Figure 3. Hardware design and components of the wearable system ObjectFinder. It incorporates a stereo camera to capture visual information about the user's surroundings, a pair of buttons, and a microphone to collect the user's commands. Simultaneously, it executes algorithms through a lightweight processor. To provide a comprehensive and immersive experience, the system delivers spatial-aware informational feedback directly to the user via bone-conduction headphones integrated into the smart glasses.

buttons that are programmed for target confirmation and function selection. A microphone is attached to the collar for audio input: participants simply speak commands to specify targets or ask questions after triggering open questions (**G1, G2**). This version of the ObjectFinder is only an initial prototype. We aim to further integrate all hardware components more compactly to improve the user's daily experience in real-world use, e.g., with smart glasses similar to Ray-Ban's Meta Wayfarer coupled to a smartphone (Waisberg et al., 2024).

3.4. Function implementation

Based on the envisioned scenario described in Section 3.1, we decomposed the object search task into five functions: object detection (F1), localization (F2), route planning (F3), scene description (F4), and open questions (F5) (**G2**). To integrate these five functions into a flexible, intent-driven object search pipeline, we define three modules: the user specifies the target, the system detects the target, and the system generates feedback, see Figure 4 (**G2**). In this pipeline, two models serve as the main functional components: an open-vocabulary object detector (YOLO-World-Large (Cheng et al., 2024)), which accepts flexible natural-language queries and detects the target object, and a GPT-4 (Achiam et al., 2023), which generates targeted scene context around the detected object via API (**G1, G3**). Both models are used in the zero-shot setting.

3.4.1. Module 1: User specifies target

The first module, *User Specifies Target*, is demonstrated in Figure 4a. When the system is turned on, the smart glasses automatically capture the frame of the scenario. GPT-4 then generates a list of objects based on the frame to initialize YOLO-World (Cheng et al., 2024). To enable flexible target queries (**G1**), the user specifies the target object using the command frame, *"Find < target >."* After receiving the command, the system will repeat the target object: *"You want to find < target >, please confirm."* For confirmation, the user should press one of the buttons on the waist bag, while the other button is for respecification. Speech-to-text is processed by Google Speech Recognition API, and text-to-speech is handled by the pyttsx3.

The relationship between the specified target objects and the initialized object list in YOLO-World falls into three types: *match*, *related to*, and *unrelated to*, as shown in Figure 5.

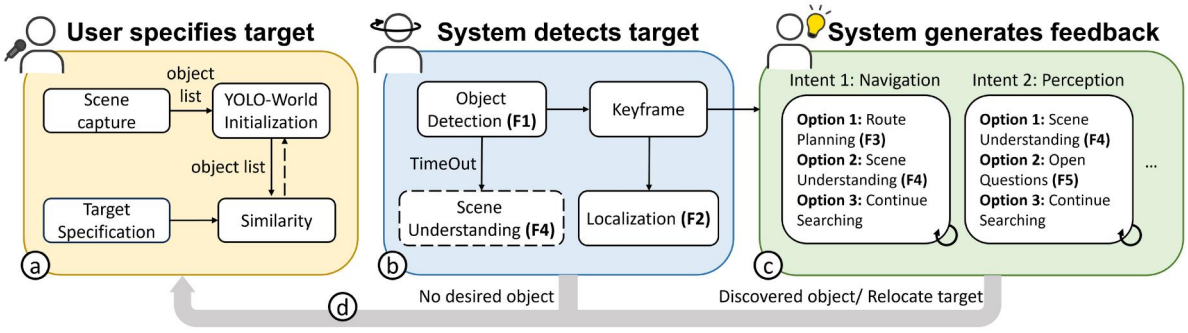

Figure 4. ObjectFinder system architecture integrates five functions into three modules for interactive object search. (a) Initially, an open-vocabulary object detector, e.g., YOLO-World, is initialized with a list of objects extracted from a scenario capture, allowing the user to identify a target object. If the target is not on the list, the object detector is reinitialized. (b) The user scans the environment. If the target is detected, localization information is provided in real-time. If not, the user can trigger scene understanding to identify what exists in the scenario. (c) The user may activate a subbranch to obtain further information based on their intent using a multimodal large language model. (d) If the user discovers other objects of interest or becomes disoriented, they can reorient themselves to locate the target.

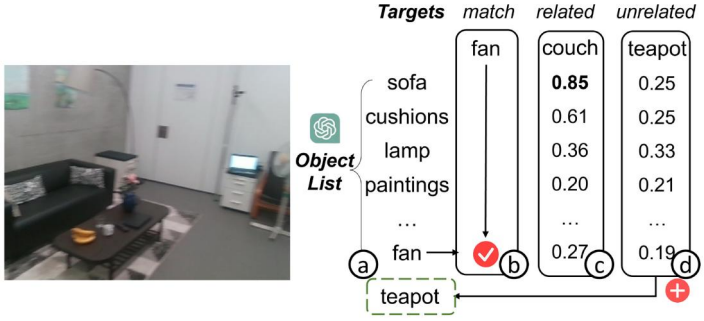

Figure 5. Initialization with target identification: (a) the list of detectable objects in YOLO-World is initialized with the first capture of the scenario. The target objects can be categorized into three types: (b) match, where the object matches an item in the list; (c) related, where the object is related to one item in the list (e.g., "*couch*" is related to "*sofa*" with 0.85 similarity); and (d) unrelated, where the object does not relate to any item in the list. In cases where the object is unrelated, the list is updated by adding the target to it.

We define *match* as the case where an object list entry's name appears as a substring of the target query. This is important because recorded speech can sometimes be unclear due to the co-designer rephrasing, such as *"Find the chair, no, office chair."* and the surrounding conversations.

For other cases, we tokenize the specified target query and each YOLO-World object class, and embed them using all-MiniLM-L6-v2 (Reimers & Gurevych, 2021). The cosine similarity between the embeddings of the target object and each class is calculated. If the maximum similarity is at least 0.8, a threshold chosen empirically, the target object is deemed to be *related to* the most similar class in YOLO-World. The mapped class is then used for subsequent detection, avoiding re-initialization. Otherwise, the target object is considered *unrelated to* current object list of YOLO-World. In this case, YOLO-World should be re-initialized with the object list updated to include the new target object. During the YOLO-World initialization process, a 3 Hz beep is played in the background to assure the user that the system is still operating (**G2**).

3.4.2. Module 2: System detects target

After YOLO-World is initialized with the target object, the user will hear an earcon to signal the start of scanning (Figure 4b). The system successfully detects the target (F1) when the confidence level of its bounding box exceeds the empirically set threshold of 0.3. This threshold was calibrated by pre-testing all target objects in both study environments to eliminate false positives, which are particularly detrimental to user trust (Hong & Kacorri, 2024b). The system captures a keyframe that includes both RGB and depth information. At the same time, another earcon sounds, signaling the user to pause and

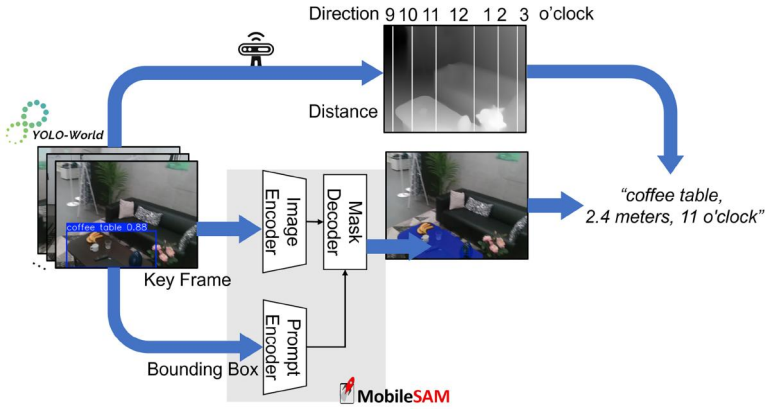

Figure 6. Object detection and localization: Each video frame is processed by YOLO-World to detect key frames in which the confidence level of the bounding box around the target object exceeds a certain threshold. Subsequently, the segmentation map generated by the bounding box, combined with the depth map, is used to provide precise localization information, including distance and direction.

orient themselves toward the target object. This frame is used to calculate localization information (F2), which is provided once in real time after the pause signal, and to query for further intent-based long-text feedback (F3–F5). If the target object is not detected within a time limit of 45 s, it is considered absent in the scenario. The user can choose to activate scene understanding (F4) or re-specify the target object (**G2**).

If the target is detected successfully, the egocentric localization information will be calculated using the keyframe and delivered in the format *Object-Distance-Direction*, as proposed by Constantinescu, Neumann, et al. (2022), as illustrated in Figure 6. When multiple objects of the same class appear within the keyframe, the egocentric information of the one with the highest confidence is provided. Egocentric information is presented in a clockwise orientation and distances in meters. Distance estimation from bounding boxes can be inaccurate (Liu et al., 2023). To improve accuracy, we added MobileSAM (Zhang et al., 2023), a compact variant of SAM (Kirillov et al., 2023), to generate segmentation masks M for YOLO World detections (Cheng et al., 2024). The distance of the object is calculated as the average depth from the key frame's depth map, masked by M.

$$Distance = \frac{\sum(\text{depth\_map} \odot \text{M})}{\sum M}. \tag{1}$$

To determine the clock direction, we use the center of the bottom edge of the frame, (x_c, y_c), as the clock center. We then compute the angle θ between the upward vertical direction and the ray from (x_c, y_c) to the bounding box center $(x_{\text{bbox}}, y_{\text{bbox}})$ as

$$\theta = \text{atan2}(x_{\text{bbox}} - x_c,\ \ y_c - y_{\text{bbox}}) \times \frac{180}{\pi}. \tag{2}$$

The resulting angle is then mapped to coarse clock-direction cues within the camera view, where 9 and 3 o'clock denote the leftmost and rightmost bounds of the camera view, respectively.

3.4.3. Module 3: System generates feedback

If the target object is detected, the system plays an earcon to prompt the user to pause and orients them toward the target. Simultaneously, it sends a keyframe from the user's egocentric view that contains the target object to the MLLM to generate long-form feedback for the user's subsequent intents, as in Figure 4c.

During refinement with the co-designer, we observed that he primarily had two intents after an object was detected. The co-designer either wanted to *navigate* to functional objects for interaction, e.g., finding a charger to charge a smartphone, or to *perceive* regions of interest, e.g., identifying items on a tabletop without interacting with them. Subsequently, the co-designer often asked follow-up

questions for additional details. Therefore, we currently implement two intent branches in this module (**G3**), with the possibility of adding more in the future.

3.4.3.1. Generating feedback based on user intent. In the process of generating intent-based feedback, the system initially uses the keyframe for the first query, while subsequent queries are based on the current egocentric frame of the user. In the *navigation* branch, the user initiates route planning (F3). After the user takes a few steps, the target object may leave the field of view. In that case, the system may no longer be able to provide updated route guidance. Consequently, the user has the option to either repeat the instruction or trigger a scene description (F4) for orientation. For example, if the user is aware that there is a desk on the route to the fan and understands from the scene description (F4) that the desk is directly in front of them, they will know *"I'm getting close to the fan."* If the user still feels lost, they can revert to the target specification to relocate the target object (**G2**). In the *perception* branch, the user can use scene description (F4) to detail the surroundings of the target object, or directly ask open questions (F5) to engage in a conversation about the target object and its surroundings over several rounds (**G2, G3**). As the co-designer always discovered objects of interest or rejected the candidate after learning about the detailed information surrounding it, the user can respecify the target object at any step with ObjectFinder (**G1, G3**).

3.4.3.2. Optimizing interaction features. Following prior work (Awad et al., 2018; Hersh, 2022; Tahoun et al., 2020), we programmed the two buttons on the waist bag to access the system's modular functions and to accommodate users' evolving intents. During refinement, we chose buttons over speech commands because speech input is susceptible to environmental noise and users may find it difficult to express evolving intent in a few sentences (Chang et al., 2024; Oumard et al., 2022). In ObjectFinder, speech input is therefore used only for target specification and for asking open-ended questions (F5), supporting efficient object search (**G1, G2**).

3.4.3.3. Enhancing feedback accessibility through prompt engineering. According to VIALM (Zhao et al., 2024), GPT-4 performs best among evaluated models for assisting users with blindness, across both human ratings (correctness, actionability, and fluency) and automated metrics. We therefore chose GPT-4 to implement the MLLM functions (F3–F5). Effective prompt engineering (OpenAI, 2023) is crucial for making MLLMs more useful and accessible to users with blindness (**G1, G3**). The response length is to the maximum within common social media alt-text limits (100–500 characters) (Perkins School for the Blind, 2024) to accommodate users' preferences for vivid responses. Notably, the co-designer scans the environment by turning his head rather than his body. When a target object comes into view, he typically pauses in his current posture, and his head and body are often misaligned. Therefore, if the system provides egocentric instructions like *"Please walk two steps forward."* or *"The desk is in front of you,"* it can lead to confusion. To resolve this ambiguity, we precede each MLLM response with the prompt *"Please align your body with the direction of your head."* This helps the system feedback accommodate different scanning strategies. For route planning, ObjectFinder provides *step*-based instructions, which are easier for users with blindness to follow (Gamage et al., 2023), and emphasizes landmarks rather than turn-by-turn directions (Jain, Teng, et al., 2023) (**G3**). The prompts are detailed in the supplementary material.

Users receive detailed information through the MLLM feedback. They may discover new target objects, lose track of the current target, or receive no detection result. In these cases, the user can trigger "continue searching" to redefine the target object, as illustrated in Figure 4d.

3.4.4. Latency

YOLO-World ran locally on the NVIDIA Jetson Nano, whereas GPT-4 was cloud-based. At the time we developed the prototype and conducted the user study, practical local deployment of open-source MLLMs was not feasible in our hardware setup. To assess the system's responsiveness, we report the latency of each module in our pipeline. Upon entering a new scene, YOLO reinitialization takes 110.9 s,

followed by 0.07 s per word for similarity calculation. Object detection takes 0.5 s, localization takes 1.1 s, and GPT response generation takes 3.5 s.

4. User study

To understand how ObjectFinder can support interactive object search in unfamiliar environments, we conducted a user study with eight participants. Specifically, we focus on the following research questions (RQs):

RQ1: To what extent does ObjectFinder provide the essential information needed for an effective object search?

RQ2: How do people with blindness perceive the targeted scene context provided by ObjectFinder's integrated design to facilitate their multifaceted object search?

RQ3: How do individuals with blindness perceive the functionality of ObjectFinder for object search compared to description-only, detection-only, and disjointly integrated systems?

RQ4: What insights from participants' interactions with ObjectFinder can guide the design of future object search systems?

4.1. Evaluation methodology

To evaluate the effectiveness, limitations, and potential of ObjectFinder, we adapt an evaluation approach used in prior work (Chang et al., 2022; Kuribayashi et al., 2023; Xie, Yu, et al., 2022). This involves comparing ObjectFinder with a strong, self-built baseline, ObjectFinder_Base, which uses closed-vocabulary object detection YOLOv8 (Varghese & Sambath, 2024). This baseline offers similar functionalities: objects can only be detected within an extensive set of predefined object classes (Lin et al., 2014) (F1), while the localization (F2) and route planning (F3) functions are available only for these detectable classes. For undetectable objects, participants can choose to trigger scene descriptions (F4) to gain an understanding of the front view. By partially decoupling detection from description, the baseline allows participants to experience the limitations of each component: unfocused descriptions **(C1)** and inflexible detection **(C2)**. This way, the baseline prototype reflects the common capabilities of existing description-only and detection-only applications such as BeMyAI and Lookout, while using identical hardware to our ObjectFinder to enable a more objective comparison.

The evaluation is carried out in two controlled indoor environments, *an office* (7.95 m^2) and *a living room* (15.96 m^2), which participants explore using a within-subjects crossover design (Jones & Kenward, 2014), as detailed in Section 4.3.

To provide retrospective context (Russell & Chi, 2014) for evaluating ObjectFinder, participants were briefly exposed to two commercial applications, BeMyAI and Lookout, which represent description-only and detection-only approaches. We considered a direct comparison out of scope due to differences in hardware, the lack of object search specificity, and the participants' prior familiarity with these applications. Instead, participants used each application to search for two objects in the living room to recall their core functionalities. This exposure refreshed their understanding of the application capabilities and supported more grounded reflections on the integrated design of ObjectFinder.

In daily life, individuals with blindness commonly rely on tactile exploration (Withagen et al., 2013) and auditory feedback (Kolarik et al., 2014) for object search, which are risky, limited in range, and unable to provide comprehensive environmental awareness. Remote sighted assistance (e.g., Aira (Aira Company, 2024), BeMyEyes (2024)) typically requires time-consuming screen sharing and frequent reorientation (Lee et al., 2020). In contrast, AI-based systems offer promising alternatives by providing real-time environmental understanding and spatial guidance. We therefore center our evaluation and discussion on AI-based systems to better understand effective computational approaches for object search assistance.

4.2. Participants and procedure

We recruited eight participants (P1–P8 in Table 2) from the local community via an existing mailing list. The participants ranged in age from 20 to 80 years ($\mu = 40.75$ years, $\sigma = 17.945$), including three women and five men. All participants were legally blind (vision $\leq 5\%$ for both eyes (World Health Organization, 2021)), with seven having acuity $\leq 2\%$. Four of them were born with blindness. For scene understanding, six of the participants had previous experience using description-based applications such as Seeing AI, BeMyAI, and Envision, while only one participant used a detection-based application, Lookout. In Table 2, we consider only the use of these applications for scene understanding, and exclude other purposes such as document reading. This study was approved by the Ethics Committee of Karlsruhe Institute of Technology (KIT), Germany. The committee issued a formal ethics approval without assigning an approval number. The video and audio recording were consented to by the participants. Participants received a compensation of € 40 for taking part in the study and were also reimbursed for their travel costs.

Each user study lasted about two hours and consisted of the following steps: (1) an introduction and tutorial of our prototype; (2) exploration of both scenarios using ObjectFinder and ObjectFinder_Base in a crossover manner (Jones & Kenward, 2014), each followed by (3) the completion of a questionnaire featuring Likert-scale items assessing function, and the NASA-TLX (Hart, 2006) for assessing cognitive load, accompanied by a short semi-structured interview; (4) short exploration of the living room scenario using the commercial applications BeMyAI and Lookout, followed by (5) another semi-structured interview in a retrospective context about the commercial applications. The interview guideline is provided in the supplementary material.

4.3. Scenario exploration

As in Section 2.1, individuals with blindness usually search for large objects as landmarks to construct mental maps of unfamiliar environments, and explore small objects on the tabletop. Thus, in each of our scenarios (living room or office), participants were asked to find six target objects: three large pieces of furniture to establish spatial orientation, followed by three smaller objects on the coffee table or desk. Although the co-designer identified the socket as a daily target object in the envisioned scenario, locating it by touch poses safety risks during the user study. Figure 7 describes a simplified procedure involving the five functions F1–F5. Table 3 specifies the target objects that need to be found in sequence and the initial MLLM functions to be triggered as the assumed initial intent when each target is detected. The layout and order of the search targets are shown in Figure 8. To our knowledge, we are the first to engineer prompts that generate route planning instructions for individuals with blindness, guiding them to objects. Therefore, our primary focus is on testing the route planning function (F3). Since the closed-vocabulary ObjectFinder_Base can only detect a limited number of target objects (Lin et al., 2014), we categorize the six target objects in each scenario into three groups: *unrelated to* (two objects), *related to* (three objects), and *exact in* (one object) COCO2017. Regarding the *related* targets, we hint participants to look for related objects in COCO2017 when they are using the baseline, but they experienced a vocabulary gap between the wording used for target specification and that in the MLLM feedback. As for the *unrelated* objects, the participants cannot even specify the target objects **(C1)**. So we provide the option for the participants to trigger the scene description function (F4) at any time to find the objects. *Cookies* are the bonus target, and we observe whether participants can discover it themselves to validate the capability of our system to prompt subsequent searches of unexpected targets **(C2)**. Figure 9 illustrates examples of how participants detected target objects and received feedback from the system during the user study. Participants navigated the environments through tactile exploration without white canes. A sighted volunteer accompanied each participant to ensure their safety throughout the sessions.

In the living room, participants briefly used Lookout (Google, 2024) and BeMyAI (2023), two commercial applications they were familiar with but that are not designed for object search, to locate a *fan* and a *teapot* and to refresh their understanding of the applications' capabilities.

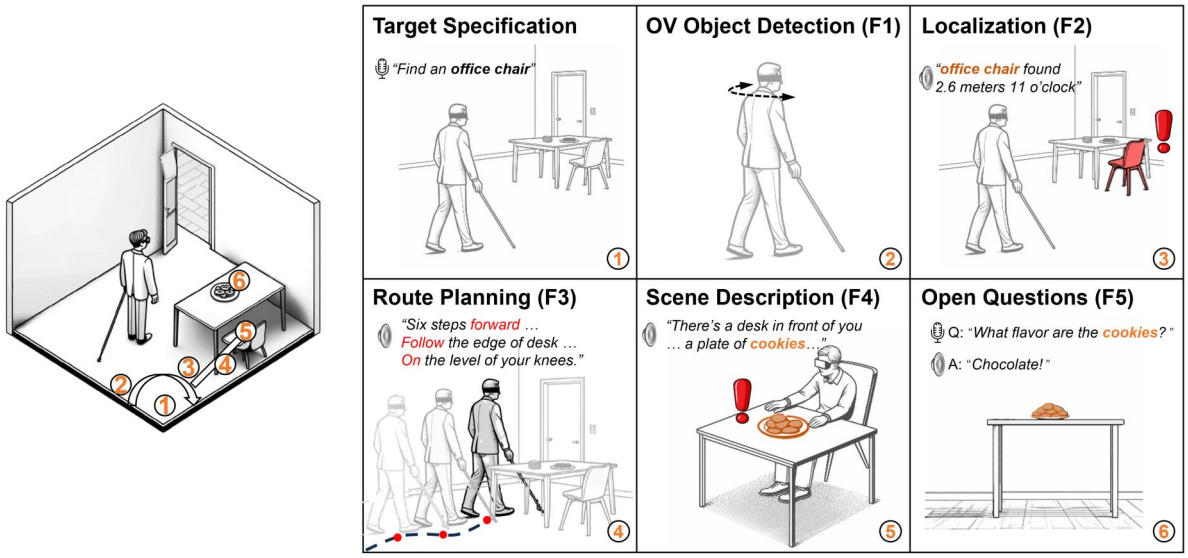

Figure 7. Simplified user study procedure focusing on two representative target objects: a large piece of furniture (an office chair) and a smaller object (a plate of cookies). The user walks into an unfamiliar environment, the office in this example. To sit in front of a desk, he or she should find an office chair first and navigate to it. Then the user sits on the office chair and defines the desk as the region of interest. The user will know what is on the desk through the scene description function, a plate of cookies in this example, and get additional information through open questions.

Table 3. Target objects in each scenario.

Scenarios		Target objects	Names in COCO (Lin et al., 2014) dataset
Office	Furniture	trash bin^{F3}, desk^{F4}, office chair^{F3}	<None>, dining table, chair
	Smaller objects	monitor^{F5}, keyboard^{F3}, headphone^{F3} (cookies^{F3})	TV, keyboard, <None> (<None>)
Living room	Furniture	fan^{F3}, coffee table^{F4}, sofa^{F3}	<None>, dining table, couch
	Smaller objects	teapot^{F3}, banana^{F3}, flower^{F5}	<None>, banana, potted plant

The superscript $\texttt{object}^{F}$ denotes the function that is activated first upon finding the target objects. F3: route planning; F4: scene description; F5: open questions.

(a) Living Room. (b) Office.

Figure 8. Layout of the two scenarios and the order of target objects. The 7th object in the office, *a plate of cookies*, is the bonus, to determine if participants are aware of its existence through the system. The photos are taken at the starting points of the task in each scenario.

4.4. Data analysis

We have both qualitative and quantitative data. For qualitative data, the user study transcripts were analyzed using the hybrid process of inductive and deductive thematic analysis proposed by Fereday and Muir-Cochrane (2006). The first author led the analysis by repeatedly reading the transcript for familiarization and coding it in multiple rounds, with questions regarding coding being resolved through discussion, and resulting themes being critically appraised within the research team. In addition to data-driven inductive coding, we applied deductive coding to interpret the system's capability to identify regions of interest **(C1)** and to facilitate the discovery of unexpected targets **(C2)**. In a

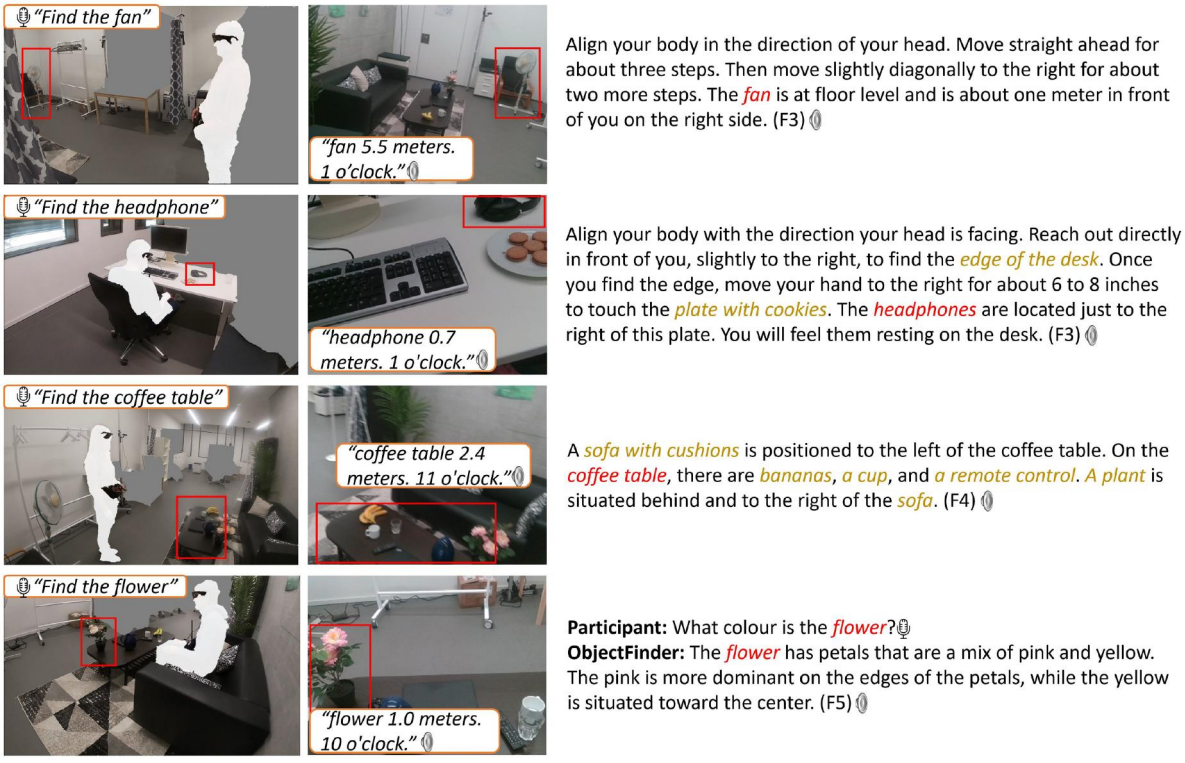

Figure 9. The first column shows examples of postures, while the second column displays key frames captured when target objects were detected (F1) along with the corresponding real-time egocentric localization information (F2). The third column presents system feedback generated by the MLLM, which uses route planning (F3) to reach both large and small items, employs scene description (F4) to describe the coffee table, and utilizes open questions (F5) to gather additional details.

workshop, the research team assigned 243 data points to 69 codes, which were further refined to 12 codes, and finally, four themes were crafted and are presented in Section 5. For quantitative data, we report only descriptive statistics (Prem, 1995) due to the small size of the user group.

5. Findings

5.1. RQ1: To what extent does ObjectFinder deliver the necessary information for effective object search?

Participants found that ObjectFinder provides an adequate amount of information, including the obligatory egocentric (distance and direction) and allocentric (relationships between the target and its surrounding objects) information, for object search. On the other hand, participants have varying perceptions of the optional information (e.g., color and alert information).

5.1.1. Amount of information

The average system feedback length for route planning (F3) and scene description (F4) was 62.90 words, with a standard deviation of 19.11. System feedback for open questions (F5) is relatively shorter. When asked whether they preferred more or less information, three participants initially described the amount as adequate. After being asked to choose between the two options, four preferred more information, while four preferred less. We noted that preferences for information quantity relate to individual differences in processing information: Some participants could ignore excess information, others felt overwhelmed. Those who preferred more information generally wanted as much information as possible. As one participant noted, *"as much as you can get. It is completely blank for me, so the more I have, the better."* (P5). P4 expressed that *"things that don't interest me, I can just ignore."* Conversely, P6 and P7 preferred less information because they were not accustomed to processing such large

amounts of visual information and found it overwhelming. P8 suggested that the information provided could perhaps be reduced after getting the first overview: *"it could maybe be a little bit less, so if you know once there are some objects on the table, then you don't really need this information the second or third time unless you ask the system what is on the table."* Additionally, we observed that the necessary amount of information varies across system functionalities. BeMyAI, used solely for scene description, received praise from three participants for its *"detailed"* information, although one deemed it excessive. In contrast, the participants noted that our ObjectFinder provided more information than necessary for efficient route planning.

5.1.2. Obligatory information

Half of the participants commented that the egocentric localization information, including both the distance and direction of the target object, was helpful for object search. We found that the participants preferred clear and intuitive distance interpretation. P1 highlighted the usefulness of different distance measures, noting, *"I think the steps are good for like when it's really near, so it's just a few steps. But if there are longer distances. I think sometimes meters might be more useful […] So it sometimes has like 0.1 meter […] you could have said 'right in front of you' or 'one step' like that."* Participants noted that the distances reported by the system often seemed larger than they experienced while approaching the object. Upon reviewing the video, we observed that participants interpreted the system's head-to-object distance as a horizontal distance. This inaccuracy may be more obvious in our small indoor settings, where larger pitch angles are more common. The left–right direction in the system feedback was also sometimes inaccurate. P5 suggested to *"specify a 20 centimeters margin,"* noting that if an object is 10 centimeters to the left or right, it's still considered in the front.

Most participants also mentioned that contextual information around the target was helpful for object search. This includes allocentric relationships among the target and nearby objects, as well as egocentric localization of those surrounding objects. For example, P3 noted, *"it's good that the objects around it are announced, so you have some idea where the object you want to find is in relation to other objects."* P8 added that such descriptions should also include distance, *"when you make the description, it holds everything, but it does not really talk about the distance."*

5.1.3. Optional information

Participants expressed mixed feelings about the relevance of color and alert information, highlighting the importance of personalizing the information to individual needs. For example, P5 and P6 appreciated the unsolicited color details. *"It told me the color of the remote control without my asking […] it helps me to visualize the environment around."* (P5). In contrast, P2 criticized excessive information, remarking, *"I want to walk from A to B, and I don't want a literature presentation of the color and the landscape description which can fill books."* Regarding alert information, P2 valued the directness of certain cues: *"what I liked was the description. It directly tells people to be careful 'move your arms', and I think it was a very clear description."* However, P3 found it superfluous, commenting, *"about the whole flavor text about 'not bumping into an object', 'sitting down', 'carefully turning the chair toward me first'. This is all not really necessary."*

5.2. RQ2: How do people with blindness perceive the targeted scene context provided by ObjectFinder's integrated design to facilitate their multifaceted object search?

Participants perceived the targeted scene context as helpful not only during the search process but also in supporting their post-search intents.

5.2.1. During search

During the object search process, participants oriented themselves and explored search options through the targeted scene context. As P1 noted, *"I was very positively surprised at […] how easy it is […] to find myself in the scene."* The value of targeted scene context for orientation is also evident by contrast: when exploring ObjectFinder_Base, P2 remarked, *"I have a description, but I have no idea where I should start searching the fan."* **(C1)**. P6 highlighted the benefits of discovering additional options **(C2)**

and understanding their arrangement, *"I didn't know there was another chair there. It gives more information about objects and their arrangement."* Apart from the targeted scene context during object search, we observed that the challenge of unknown-unknowns (Bigham et al., 2017) for search options in unfamiliar environments persists. This can be mitigated by providing an initial overview of the environment: *"I could imagine if you really don't know the room, and you want to first [have an] overview of what is in the room, it was pretty detailed."* (P8).

5.2.2. After search

In addition, the targeted scene context also facilitated post-search intents, such as navigation and perception of the region of interest, while reducing the need for physical exploration. For navigation, P1 noted when searching for the fan with ObjectFinder: *"I think it was very good [to know] where I am. When it said where my target object was, it also told me the surrounding objects, because maybe if I find the surrounding object first, then I know, OK, I'm close."* P6 particularly valued the feature, noting: *"if it's a new desk, I don't have to feel around to know where the computer is. I don't have to move far right or left to find the position easily."* For perception, four participants mentioned that the system's detailed output was helpful for discovering objects on the tabletop **(C2)**, "*When I first got into the scene, I got a very detailed description of the coffee table […] so I knew what I could expect to find there."* (P1)

5.3. RQ3: How do individuals with blindness perceive the functionality of ObjectFinder for object search compared to description-only, detection-only, and disjointly integrated systems?

Participants appreciated the ability to actively specify targets and the provision of comprehensive egocentric and allocentric information for object search, in contrast to description- and detection-only applications. Despite having similar functions, participants observed a significant performance drop with the disjointly integrated system, due to untargeted descriptions and limited functionality for navigation.

5.3.1. Reflections on ObjectFinder in contrast to commercial systems

Regarding the differences between the commercial applications and the prototypes, participants valued that they could actively define the target objects with ObjectFinder, making it more reliable in searching for specific items. As P6 mentioned, *"what I also liked about this system is speaking to it, I find that more targeted."* P3 echoed this sentiment, highlighting the biggest advantage: *"The biggest advantage […] contrary to BeMyAI is that you can specify what you want to find."* **(C1)**. Furthermore, as previously analyzed, ObjectFinder delivers essential information for object search: egocentric localization (distance and direction) and allocentric relationships between the target and surrounding objects. These cues are not typically supported by description- or detection-only systems. For example, after using BeMyAI, P8 mentioned, *"BeMyAI had no information regarding the distances, so you have […] a less sense of location."* The lack of localization information (P2, P8) and the absence of context regarding object relationships (P2) were mentioned as reasons why participants did not prefer Lookout, highlighting the potential for solutions that can combine multiple aspects of object search.

5.3.2. Workload differences between integrated and disjoint systems

The first two columns in Table 4 show workload ratings for object search tasks with ObjectFinder and the self-built baseline ObjectFinder_Base, which implements description and detection as disjoint components. Participants reported lower workload with ObjectFinder than with the baseline, especially in the performance dimension ($M = 4.69$, $SD = 2.02$ vs. $M = 7.25$, $SD = 2.38$). Six participants rated ObjectFinder lower than the baseline on the performance dimension, one rated the two systems equally, and one rated ObjectFinder higher. Mental demand and effort dimensions showed considerable variability among subjects ($SDs \sim 2 - 3.5$), while temporal demand was low for both systems ($M = 2.63$). These quantitative results suggest that ObjectFinder's integrated design led to better performance under low time pressure. Regarding mental and temporal demand, P8 expressed, *"The second time (ObjectFinder_Base), I was a bit more frustrated that he did not really find [specific things] or he did take much longer than the first example (ObjectFinder)."* Other participants attributed their ratings to

Table 4. Assessment of workload required to complete tasks with two systems (ObjectFinder *vs.* ObjectFinder_base) and in two scenarios (office *vs.* living room) using NASA-TLX. Bold values indicate lower (better) workload scores.

	Systems				Scenarios			
	ObjectFinder		ObjectFinder_base		Office		Living room	
Sub-scale	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Mental demand	**4.25**	2.96	5.38	3.42	**4.50**	2.93	5.13	3.52
Physical demand	**2.44**	1.76	3.00	2.39	**2.69**	2.25	2.75	1.98
Temporal	**2.63**	2.00	**2.63**	1.30	**2.13**	1.46	3.13	1.73
Performance	**4.69**	2.02	7.25	2.38	**4.94**	2.43	7.00	2.27
Effort	**4.13**	2.03	5.38	3.11	**4.75**	3.11	**4.75**	2.25
Frustration	**4.13**	3.48	4.38	3.07	**4.00**	3.46	4.50	3.07

Scores represent workload assessed across both systems and both scenarios separately. Higher scores indicate higher workload (worse), except for performance where higher scores indicate poorer self-assessed performance.

the difference between ObjectFinder and ObjectFinder_Base in identifying target objects and providing targeted information. As P5 noted: *"the second system (ObjectFinder_Base) was not able to identify products easily. I needed to trigger scene description to understand the environment a lot, as opposed to the first system (ObjectFinder)."* P1 noted that the functional limitations stem from the disjoint implementation of description and detection components, *"I couldn't just get the navigation description, but had to go find it myself."*

5.3.3. Task performance differences between integrated and disjoint systems

Though not statistically significant with eight participants, we report task performance metrics for reference. Each participant searched for six target objects with each system, yielding 48 target-object search trials per system (8×6). ObjectFinder achieved 47/48 successes, versus 43/48 for ObjectFinder_Base. With ObjectFinder_Base, four participants (P3, P5, P6, P7) abandoned one search after the target object was repeatedly not mentioned in the descriptions (F4), while P1 gave up searching for the headphones, believing they were not present based on earlier descriptions. P2 abandoned the search for the trash bin with ObjectFinder after becoming disoriented. ObjectFinder successfully detected all target objects in the user study, while the detection accuracy of ObjectFinder_Base was 33% across all trials. Although the theoretical accuracy of ObjectFinder_Base is 66.7% according to Table 3, participants often made multiple attempts when objects were not detected, which lowered the overall accuracy. The average number of trials per object was lower for ObjectFinder ($M = 1.17$, $SD = 0.37$) than for ObjectFinder_Base ($M = 1.63$, $SD = 1.13$). Similarly, for trials requiring route planning (F3), task completion time was shorter with ObjectFinder ($M = 70.97$ s, $SD = 41.57$) than with ObjectFinder_Base ($M = 79.41$ s, $SD = 55.15$).

5.3.4. Scenario effects on workload and system limitations

Regarding the impact of scenarios in Table 4, the temporal demand was in the living room ($M = 3.13$, $SD = 1.73$) compared to the office ($M = 2.16$, $SD = 1.46$), with four participants rating them equally and four rating the living room higher. Two participants rated performance in the living room more than 4 points higher than in the office, and upon review, these participants used ObjectFinder_Base for exploration. In the living room, overlooking the *coffee table* as an obstacle on the way to the *sofa* was mentioned by four participants and may have made the scenario more challenging. For example, P1 noted: *"[…] the possibility of overlooking some obstacles like the coffee table […] if I hadn't known that there was the table, I would have just run into it, which is not that nice."* Retrieving keyframes for MLLM information reveals that the landscape orientation of the camera at eye level provides a limited vertical field of view, and participants were not accustomed to lowering their heads to detect obstacles. Although participants achieved the same task completion rate (45/48) in both settings, they took longer to complete each object-search task in the living room ($M = 83.22$ s, $SD = 59.17$) than in the office ($M = 67.09$ s, $SD = 41.34$).

Although ObjectFinder was evaluated in controlled indoor scenarios, P2 envisioned broader use cases, such as *"[…] why not outdoor? For example, in a park to find a bench […] in a hotel […] to find an information desk,"* which are often challenging due to limitations in tactile exploration.

5.4. RQ4: What insights from participants' interactions with ObjectFinder can guide the design of future object search systems?

Participants suggested that interaction features and hardware for object search should support personalization. Additionally, they considered object detection, localization, and route planning to be important functions of object search.

5.4.1. Interaction features

Participants highlighted the need for efficient and flexible interaction for both input and receiving information. Regarding target specification, P8 described the voice commands as *"pretty, pretty easy,"* while P4 suggested *"maybe it's better to have the ability to switch between a list of maybe recognized objects and voice commands."* P5 mentioned that she was not yet accustomed to the earcon indicating when to stand still. Besides the current earcons, P6 suggested replacing the current spoken option confirmation with an additional earcon.

As mentioned before, some participants preferred more information, while they suggested *"implementing a skip option"* (P1) and the ability to *"switch information on and off optionally"* (P2). P4 suggested adding a main menu to easily switch between object search (detection and navigation) and scene description. The importance of efficient information acquisition was also reflected in participants' references to commercial applications. P1 highlighted the interaction feature of Lookout, *"I don't have to take a picture and wait for a response."* P2 shared his experience that receiving explanations from a sighted person on BeMyEyes *"works better because it's without delay."*

5.4.2. Software requirements

We asked participants to rate the functions (F1–F5) before and after exploring ObjectFinder and ObjectFinder_Base, in terms of their importance and level of interest for object search, as captured by Likert scores (Figure 10). Among the five functions, localization information (F2) was rated as the most important. The functions of object detection (F1), localization information (F2), and route planning (F3) are important for the object search task ($\mu > 4.0$), with reduced standard errors after exploration. However, two of these functions (F2 and F3) are not available in description-only or detection-only commercial applications, as they rely on the integration of both components. P8 mentioned the advantage of ObjectFinder is *"the distance and also the guiding function which is not really available for the other both apps."*

Additionally, we note that the software should make few mistakes. Half of the participants reported disliking Lookout's frequent misidentifications. As P5 noted, *"it misidentified objects […], like there's no dishwasher,"* and at times, the *teapot* was mislabeled as a *helmet* or *mouse*.

5.4.3. Hardware requirements

Participants noted both advantages and disadvantages of capturing scenarios with glasses compared to a cellphone. Three participants experienced reduced mental effort in determining the camera's orientation when using glasses that capture the egocentric view. *"It's always harder for me to think exactly about what the phone is capturing with the phone camera, I found that better with the glasses."* (P6).

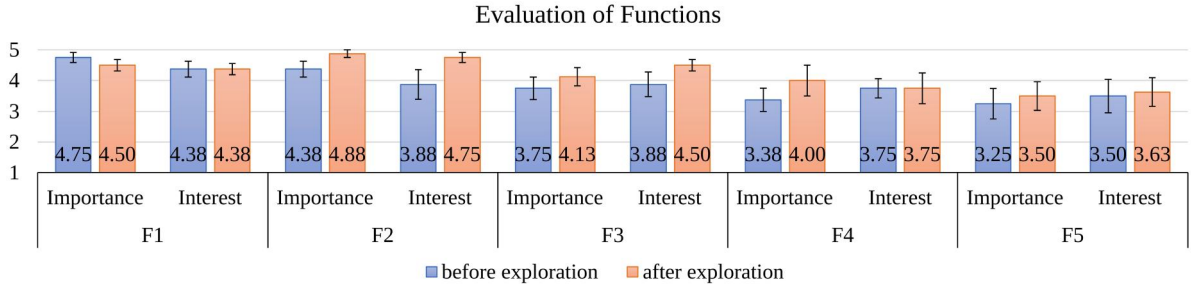

Figure 10. Evaluation of the five functions in terms of importance and level of interest, using 5-point Likert scale (1 = not at all, 5 = very much), before and after exploring two scenarios. Error bars represent standard. F1: object detection; F2: localization information; F3: route planning; F4: scene description; F5: open questions.

On the contrary, users required external cues to lower their heads to capture objects lower in the scene. Two participants who explored the office using ObjectFinder_Base failed to capture items on the desk without lowering their heads and remained unaware of a plate of *cookies* that was also present. **(C2)** Additionally, using glasses poses a challenge, as they represent an extra item to carry and can be forgotten alongside smartphones (P4, P5). The price of the glasses was also mentioned to be considered.

6. Technical evaluation of MLLM outputs

To assess the reliability of the MLLM-generated feedback (Li et al., 2026), we conducted a post-hoc technical evaluation of all MLLM outputs collected during the user study and the tutorial before. In addition to GPT-4V (Achiam et al., 2023), which was deployed in the original system, we re-ran the same keyframes and prompts through two recent models, Claude Opus 4 (Anthropic, 2025) and Gemini 2.5 Flash (Gemini Team, Google, 2025), to examine whether advances in multimodal reasoning mitigate the spatial limitations observed in the original outputs.

6.1. Evaluation protocol

Two independent annotators with knowledge of indoor spatial layouts reviewed all 169 interactions (62 route planning, 84 scene description, 23 open question) across three models, yielding 507 annotated responses in total. Each annotator worked with anonymized model labels whose assignment was randomized per interaction to prevent bias. The annotation scheme was designed per function type:

- Route planning (F3): Each response was judged on three binary dimensions: *direction correctness* (whether directional instructions match the actual scene), *collision safety* (whether following the instructions would avoid obstacles), and *actionability* (whether instructions are specific enough for a blind user to execute).
- Scene description (F4): Annotators extracted individual spatial claims from each response (*e.g., "a sofa is on the left"*) and labeled each claim for *object existence* (present vs. hallucinated) and *spatial correctness* (correct, incorrect, or ambiguous).
- Open questions (F5): Each response was rated for *factual correctness* (correct, partially correct, or incorrect) with respect to the visual content of the keyframe.

6.2. Route planning results

Table 5 summarizes the route planning evaluation. GPT-4V, the model used in the original user study, achieved a relatively low direction correctness rate of 79.8% and a collision safety rate of 77.4%. Both newer models showed substantial improvements, with Gemini 2.5 Flash reaching 92.7% direction correctness and both Claude Opus 4 and Gemini 2.5 Flash achieving 87.9% safety. Actionability followed a similar pattern, rising from 82.3% (GPT-4V) to 89.5% (Gemini 2.5 Flash).

6.3. Scene description results

Table 6 presents the claim-level evaluation of scene descriptions. Across both annotators, GPT-4V produced an average of 3.6 spatial claims per response with a hallucination rate of 10.7% and a spatial accuracy of 82.5%. The newer models showed marked improvements on all metrics. Gemini 2.5 Flash achieved the lowest hallucination rate (1.8%), while Claude Opus 4 attained the highest spatial accuracy (96.3%) and generated the most detailed descriptions (5.0 claims per response on average).

Table 5. Two independent annotators assessed 62 interactions per model across three dimensions. Bold values indicate better performance, *i.e.*, higher directional correctness, higher collision safety, and better instruction actionability.

Model	Direction ↑	Safety ↑	Actionability ↑
GPT-4V (Achiam et al., 2023) (original)	79.8	77.4	82.3
Claude Opus 4 (Anthropic, 2025)	91.9	**87.9**	86.3
Gemini 2.5 Flash (Gemini Team, Google, 2025)	**92.7**	**87.9**	**89.5**

Values represent the averaged positive rate (%) across both annotators.

Table 6. Claim-level evaluation of scene description outputs (F4). Bold values indicate better performance, *i.e.*, lower hallucination rate and higher spatial accuracy.

Model	Claims	Claims/resp.	Halluc. ↓	Spatial acc. ↑
GPT-4V (Achiam et al., 2023) (original)	302	3.6	10.7	82.5
Claude Opus 4 (Anthropic, 2025)	417	5.0	3.5	**96.3**
Gemini 2.5 Flash (Gemini Team, Google, 2025)	387	4.6	**1.8**	93.8

Two annotators independently extracted and assessed spatial claims from 84 descriptions per model. Hallucination rate measures the percentage of claims referring to non-existent objects.

6.4. Open question results

For open questions, factual correctness improved from 80.4% with GPT-4V to 89.1% with both Claude Opus 4 and Gemini 2.5 Flash. The small sample size (23 interactions per model) limits the granularity of this comparison, but the trend is consistent with the improvements observed for the other two functions.

These results confirm quantitatively the spatial reasoning errors identified in Figure 9: GPT-4V's route instructions carried a 22.6% collision risk and its scene descriptions contained 10.7% hallucinated objects. Newer models substantially reduce but do not eliminate these errors, with the best-performing model still producing a 12.1% collision risk and a 1.8% hallucination rate. These residual errors suggest that no single information source should be trusted in isolation. Instead, the sensor-based egocentric localization (F1–F2) and the MLLM-generated contextual feedback (F3–F5) serve as complementary channels: users can cross-reference low-level spatial cues (e.g., distance and clock direction from depth sensing) against high-level scene context to form a more robust mental model of the environment, much as a sighted companion offers both precise directions and broader contextual narration.

7. Discussion

Our research highlights that object-search systems have potential as an assistive technology for people who are blind. Here, we want to discuss future directions for the development of such systems, outlining challenges and opportunities for the Human–Computer Interaction (HCI) and accessibility research communities.

7.1. Enhancing interaction and information design for object-search systems

In our study, we uncovered a range of features and characteristics of ObjectFinder that offer interesting avenues for future work, either through iteration on our system or integration into comparable systems. This subsection focuses on interaction flow, context delivery, and intent handling.

7.1.1. Adapting seamless context delivery to dynamic scenarios and varying use cases

In our study, participants acknowledged that ObjectFinder effectively provides allocentric and egocentric information as targeted scene context for object search, yet they still expressed a desire for more seamless context: they wanted a brief overview of the scene before starting a search, as well as a simple main menu to choose between scene description and object search functions. While WorldScribe (Chang et al., 2024) is not a goal-oriented system, its ability to provide scene descriptions with dynamic latency and granularity could complement ObjectFinder's goal-oriented capabilities. Future work could therefore investigate how to unify general-purpose exploration and goal-oriented object search, for example by introducing an explicit exploration mode and a goal mode, and allowing users to switch between them through lightweight controls (e.g., a dedicated button or short voice commands such as *"help me find… "* to enter goal mode, and *"stop searching"* to return to exploration). Mode clarity could be enhanced through distinct earcons and consistent phrasing, building on customizable audio descriptions (Cheema et al., 2025; Natalie et al., 2024) and earcon-based mode awareness (Brewster et al., 1993; Monsaingeon et al., 2023).

7.1.2. Providing tailored and adaptive information for object search

Compared to systems based solely on descriptions or detection, and to the disjointed ObjectFinder_Base, participants perceived that ObjectFinder's seamlessly integrated design delivers essential information for object search, including egocentric cues (distance and direction to the target) and allocentric cues (relationships between target and surrounding objects), and enables crucial functions (object detection, localization, and route planning). However, participants expressed mixed views about the amount of information and about optional details, e.g., color and alerts. Future object-search systems should allow users to adjust and personalize both the amount and type of information, e.g., by skipping verbose content and personalizing descriptive attributes.

Participants also suggested how information delivery could be improved. Participants wanted distances not only to the target but also to surrounding objects, and preferred less repetition as they became familiar with a space. Future designs could therefore combine speech with spatial audio (Afonso-Jaco & Katz, 2022): speech feedback offers initial clarity when a user first encounters an object or scene, while spatialized earcons could provide continuous, low-cognitive-load tracking during active search and navigation (Blum et al., 2012; Katz et al., 2012). For instance, a spatial earcon mapped to the target's location could update in real time as the user moves, with pitch or tempo modulations conveying proximity, as demonstrated by systems such as Microsoft Soundscape (Microsoft Research, 2018). As users build a mental model of the space, the system could gradually shift from speech-dominant to earcon-dominant feedback to further reduce cognitive load.

7.1.3. Facilitating intent recognition in object search

Although LLMs can infer user intent as an auxiliary cognitive system, prior work shows that users with blindness struggle to articulate their complex intents in just a few sentences and often refine their goals during exploration (Chang et al., 2024). To address this, ObjectFinder offers intent selection via a set of refined prompts that users trigger with buttons, explicitly indicating their current intent and updating it with new triggers along the path. This design makes the interaction more controllable. However, since the current version of ObjectFinder is dedicated to unfamiliar scenarios, the options are predefined and relatively general, and not yet programmable (Herskovitz et al., 2024), which limits adaptation to users' habitual, idiosyncratic intents (e.g., *"find the laptop on my workstation"*). Future systems could either make these intent options programmable, allowing users to define reusable intent programs tailored to their own routines (Herskovitz et al., 2024; Wen et al., 2024), or incorporate online and continual personalization methods for LLMs that incrementally adapt to individual histories and preferences (Chen et al., 2024; Liu et al., 2025; Shi et al., 2026; Tan et al., 2024; Woźniak et al., 2024).

7.2. Strengthening robustness and real-world integration

This subsection focuses on practical deployment considerations, and how ObjectFinder can fit into established tools and strategies of users with blindness.

7.2.1. Improving system reliability and integration with other assistive technology

Unsurprisingly, participants expressed a preference for a system that is error-free and can accurately locate searched objects. The egocentric distance to the target should be measured horizontally instead of the current head-to-object distance d. With the pitch angle θ from the IMU integrated in most smart devices, the horizontal distance can be approximated as $d \cdot \cos\theta$. Likewise, we noted instances when furniture fell outside the camera's field of view. For example, living room furnishings were lower than those in the office in our laboratory setting, and the target coffee table was always missed. Incorporating additional sensing, like LiDAR (Liu et al., 2021) and omnidirectional camera (Kawaharazuka et al., 2024), could enhance scene perception over larger areas, reducing the need for physical scanning efforts. Wider FoV cameras, however, may exceed the coverage of depth sensors. In such cases, monocular depth estimation methods (Piccinelli et al., 2024; Yang, Kang, et al., 2024) could serve as an alternative for distance inference beyond the depth sensor's range. In terms of software, once the infrastructure is upgraded to deploy all models on a remote server, SLAM (Matsuki et al.,

2024; Yan et al., 2024; Zhu et al., 2024) and AI agents (LLMs with memory) (Xu et al., 2025; Yang, Yang, et al., 2025, Anthropic, 2025) can support holistic scene understanding.

The system should be lightweight and offer intuitive interaction features. Earcons should be learnable, and a training session to acquaint users with the meanings of earcons at the outset is advisable. According to Jakob's Law (Jakob, 1999), the design of object search systems should incorporate the use of daily tools of users with blindness, e.g., using cell phones to replace Jetson Nano for data processing.

7.2.2. Augmenting users' established strategies with ObjectFinder

People with blindness rely on established orientation and mobility strategies to find objects, including cane (Scott, 2025), maintaining organized environments with consistent object placement and tactile labels (Bilal et al., 2025; Macular Disease Foundation Australia, n.d), and using systematic search patterns grounded in mental maps and salient landmarks (Guide Dogs for the Blind Association, 2025; Schinazi et al., 2016). Outdoors, they extend these strategies into route and survey knowledge (Ottink et al., 2022) anchored by *"crucial objects"* (Islam et al., 2024) such as crossings and traffic lights that structure exploration and help confirm their location. While active exploration remains central, participants in our study reported that ObjectFinder can ease this process by making them aware of target candidates and tabletop items without extensive physical exploration, and by providing targeted scene context for orientation, although they still rely on their hands to locate the objects. This suggests that ObjectFinder augments, rather than replaces, tactile exploration. ObjectFinder is designed as a wearable to keep users' hands free for tactile exploration. Future work should investigate which kinds of contextual knowledge are most crucial for supporting these established strategies in different scenarios, and how mobility training and users' strategies might evolve when they use such assistive systems.

7.3. Tensions and concerns regarding vision- and AI-based assistive technology

There are tensions and concerns that need to be resolved for systems such as ObjectFinder to effectively and safely support object searching.

7.3.1. User habits and needs regarding lighting conditions for camera-based systems

Although computer vision can compensate for vision loss, good lighting is crucial for camera-based systems. However, this conflicts with the fact that light plays a different role in the lives of people with blindness: Persons who are legally blind may not use light in their homes and workplaces in the same way as sighted developers would anticipate (Gonzalez Penuela et al., 2024), and people who do have residual vision may not find lighting conditions required by camera systems comfortable. Considering such scenarios, future systems could explore integrating thermal imaging as an additional modality to complement RGB and depth, reducing dependence on bright, potentially uncomfortable lighting conditions.

7.3.2. Addressing safety concerns in the context of AI

The failure to detect low pieces of furniture can be attributed not only to technical constraints, but also to scanning strategy limitations, e.g., participants may not anticipate the height of certain furniture and therefore fail to direct their scanning downward to locate it. While established gaze-based environmental scanning strategies exist for people who are blind or have low vision (Riddering, 2023), no specific scanning strategies for smart glasses tailored to individuals with blindness have been developed. With advancements in wearable systems, integrating these techniques into mobility training is now both viable and beneficial.

Despite new technological solutions (which may introduce additional challenges), this case illustrates a safety tension: Users are encouraged to rely on such systems, yet vision- and AI-based tools can be unreliable, e.g., in assistive technology (Alharbi et al., 2024). This gap between promise and reliability calls for transparent communication and collaboratively negotiated expectations with users who are blind, for instance by exposing calibrated reliability cues (such as confidence-linked phrasing) and co-designed fallbacks to established non-technological strategies.

7.3.3. Understanding the limitations of technology for object search

Finally, the HCI and accessibility communities have previously questioned technologies that narrowly equate independence with individual self-sufficiency. Beatrice (2021) critically appraised navigation technologies for people with blindness, arguing that in some situations collaboration with others may be more appropriate than a technological solution. Building on the principle of interdependence (Bennett et al., 2018), assistive technology should account for the social context and collective access, raising questions such as: *Is it really necessary for a person with blindness to rely on ObjectFinder at work, or could non-disabled colleagues instead make a greater effort not to misplace or alter their desk?* Here, we would like to make clear that we envision systems such as ObjectFinder tools that can be leveraged by people with blindness in their day-to-day life, but that we do not believe that these should replace collective access labor and shared responsibility for everyday spaces.

8. Limitation and future work

There are a few limitations that need to be taken into account when interpreting our findings. This research, conducted in small indoor settings of our lab, does not investigate the effectiveness of ObjectFinder in larger, real-world public, private, or outdoor environments. In particular, real-world conditions introduce challenges that our controlled lab setting did not capture, including changing lighting, moving objects causing false detections or intermittent occlusion, and crowded scenes reducing detection confidence and depth estimation. Potential mitigations include temporal filtering, adaptive confidence thresholds, and object tracking. Our work takes a first step toward interactive object search for people with blindness, and we leave evaluation under dynamic real-world conditions for future work. We therefore recruited only participants from nearby regions.

Several integration challenges remain across the components of ObjectFinder. The confidence threshold was calibrated in a controlled laboratory environment to minimize false positives. Adapting this threshold to diverse real-world scenarios remains an open challenge, as environmental variability can shift the effective detection distribution. Although ObjectFinder possesses MLLM capabilities to infer surroundings beyond the immediate view, it cannot reliably remember explored spaces. Thus, each environment remains unfamiliar to the system, limiting its ability to provide a holistic understanding to the user. Furthermore, due to the latency of MLLM-based functions, feedback is generated from frames captured several seconds earlier and may no longer reflect the user's current position, even when the user is signaled to pause. Future iterations could integrate continuous pose tracking to compensate for user movement during processing, keeping spatial feedback synchronized with the user's real-time location. Even though glasses are preferred for orientation, the ObjectFinder prototype is equipped with relatively heavy hardware, as it is designed for upstream technology development rather than for everyday use at this stage.

Our first goal was to assess whether the interactive object-search system with seamless integration of description and detection components tackles the multifaceted object search task, and to define the future direction for the development of the object search system. After addressing the challenges and implementing the suggestions from participants, we intend to conduct a larger-scale quantitative evaluation of the system in comparison with business-as-usual approaches, engaging more participants with blindness in real private and public scenarios across different cities to evaluate ObjectFinder's effectiveness in real-world conditions. Given the residual error rates observed in our technical evaluation, ObjectFinder should be used as a supplementary information source alongside existing mobility aids such as white canes or guide dogs, rather than as a standalone navigation system.

9. Conclusion

In this work, we explored the design and development of a prototype that integrates detection and description seamlessly to enable open-vocabulary interactive object search by people with blindness. With our prototype, we address limitations of existing systems that are either description- or detection-only: locating regions of interest and discovering incidental targets. The system feedback is tailored to various user intents. Our user study suggests that this approach is promising, as it provides essential

egocentric localization, allocentric relational cues, core functionalities, intent-driven scene context, and a conversational interface. Overall, our work represents an initial step toward developing AI-based assistive technology that supports the multifaceted object search, providing first insights into user requirements and application challenges. Here, we hope that our work will encourage and facilitate further development of object-search systems, and that it will inspire future studies into the experiences that blind people have with such technologies.

Acknowledgements

We thank Kaige Wang and Zhi Wang for supporting the user study as volunteers, and Zirui Wang and Shaofang Quan for joining the MLLM output annotation. We adjusted in some cases the grammar of the article using ChatGPT.

Authors' contributions

CRediT: **Ruiping Liu**: Conceptualization, Data curation, Formal analysis, Investigation, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Jiaming Zhang**: Conceptualization, Project administration, Software, Visualization; **Angela Schön**: Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review & editing; **Karin Müller**: Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Writing – review & editing; **Junwei Zheng**: Conceptualization, Software, Visualization; **Kailun Yang**: Conceptualization, Writing – review & editing; **Anhong Guo**: Formal analysis, Investigation, Visualization, Writing – review & editing; **Kathrin Gerling**: Formal analysis, Methodology, Visualization, Writing – review & editing; **Rainer Stiefelhagen**: Funding acquisition, Project administration, Resources, Supervision.

Disclosure statement

The authors report that there are no competing interests to declare.

Funding

This work was supported in part by the Ministry of Science, Research and the Arts of Baden-Wurttemberg (MWK) through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, in part by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF), and in part by Karlsruhe House of Young Scientists (KHYS). We thank HoreKA@KIT, HAICORE@KIT, and bwHPC supercomputer partitions.

ORCID

Ruiping Liu http://orcid.org/0000-0001-5245-2277
Jiaming Zhang http://orcid.org/0000-0003-3471-328X
Karin Müller http://orcid.org/0000-0003-4309-1822
Junwei Zheng http://orcid.org/0009-0005-4390-3044
Kailun Yang http://orcid.org/0000-0002-1090-667X
Anhong Guo http://orcid.org/0000-0002-4447-7818
Kathrin Gerling http://orcid.org/0000-0002-8449-6124
Rainer Stiefelhagen http://orcid.org/0000-0001-8046-4945

Data availability statement

The data that support the findings of this study are openly available in the Open Science Framework (OSF) at https://osf.io/tcraq, DOI: https://doi.org/10.17605/OSF.IO/TCRAQ.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. (2023). *GPT-4 technical report.* arXiv preprint arXiv:2303.08774.

Afonso-Jaco, A., & Katz, B. F. G. (2022). Spatial knowledge via auditory information for blind individuals: Spatial cognition studies and the use of audio-VR. *Sensors (Basel, Switzerland)*, *22*(13), 4794. PMID: 35808291; PMCID: PMC9268803. https://doi.org/10.3390/s22134794

Ahmetovic, D., Sato, D., Oh, U., Ishihara, T., Kitani, K., & Asakawa, C. (2020). *ReCog: Supporting blind people in recognizing personal objects* [Paper presentation]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (Chi, '20) (pp. 1–12). ACM. https://doi.org/10.1145/3313831.3376143

Aira Company. (2024). *Aira: Visual assistance for blind and low-vision individuals*. https://www.aira.io

Alharbi, R., Lor, P., Herskovitz, J., Schoenebeck, S., & Brewer, R. N. (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* (pp. 1–17). ACM.

Anthropic. (2025). *System card: Claude Opus 4 & Claude Sonnet 4*. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf

ATMAPS Project. (n.d.). *User Requirements and Specifications Report (Deliverable D2.1)*. Project Deliverable. ATMAPS Consortium. https://www.atmaps.eu/deliverables/ATMAPS-D_2_1-User_requirements_and_specifications_report.pdf

Awad, M., El Haddad, J., Khneisser, E., Mahmoud, T., Yaacoub, E., & Malli, M. (2018). *Intelligent eye: A mobile application for assisting blind people* [Paper presentation]. Proceedings of the IEEE Middle East and North Africa Communications Conference (MENACOMM) (pp. 1–6). IEEE. https://doi.org/10.1109/MENACOMM.2018.8371005

Aydemir, A., Pronobis, A., Gobelbecker, M., & Jensfelt, P. (2013). Active visual object search in unknown environments using uncertain semantics. *IEEE Transactions on Robotics*, *29*(4), 986–1002. https://doi.org/10.1109/TRO.2013.2256686

Bala, M. M., Vasundhara, D. N., Haritha, A., & Moorthy, C. H. V. K. N. S. N. (2023). Design, development and performance analysis of cognitive assisting aid with multi sensor fused navigation for visually impaired people. *Journal of Big Data*, *10*(1), 21. https://doi.org/10.1186/s40537-023-00689-5

Beatrice, V. (2021). AI assistive technology for extending sighted guiding. *SIGACCESS Accessible Computing*, *129*(Article 7), 5. https://doi.org/10.1145/3458055.3458062

BeMyAI. (2023). *Introducing Be My AI*. https://www.bemyeyes.com/blog/introducing-be-my-ai/.

BeMyEyes. (2024). *Be My Eyes App*. https://www.bemyeyes.com/.

Bennett, C. L., Brady, E., & Branham, S. M. (2018). *Interdependence as a frame for assistive technology research and design* [Paper presentation]. Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (Galway, Ireland) (ASSETS '18) (pp. 161–173). Association for Computing Machinery. https://doi.org/10.1145/3234695.3236348

Bigham, J. P., Jayant, C., Miller, A., White, B., & Yeh, T. (2010). *VizWiz::LocateIt – Enabling blind people to locate objects in their environment* [Paper presentation]. 2010 Computer Society Conference on Computer Vision and Pattern Recognition – Workshops (pp. 65–72). IEEE. https://doi.org/10.1109/CVPRW.2010.5543821

Bigham, J. P., Lin, I., & Savage, S. (2017). *The effects of "Not Knowing What You Don't Know" on Web Accessibility for Blind Web Users* [Paper presentation]. 19th International ACM SIGACCESS Conference on Computers and Accessibility (Baltimore, Maryland, USA) (ASSETS '17) (pp. 101–109). Association for Computing Machinery. https://doi.org/10.1145/3132525.3132533

Bilal, S., Rebate, J., Jacobs, D. M., & Sevillano, V. (2025). Hotel stays of individuals with a visual impairment: A qualitative study with a focus on sensory substitution. *Disability and Rehabilitation. Assistive Technology*, *20*(8), 2885–2901. https://doi.org/10.1080/17483107.2025.2511982

Blum, J. R., Bouchard, M., & Cooperstock, J. R. (2012). What's around me? Spatialized audio augmented reality for blind users with a smartphone. In Alessandro, P., & Tao, G. (Eds.), *Mobile and ubiquitous systems: Computing, networking, and services* (pp. 49–62). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30973-1_5

Boldu, R., Matthies, D. J. C., Zhang, H., & Nanayakkara, S. (2020). AiSee: An assistive wearable device to support visually impaired grocery shoppers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(4), 1–25. https://doi.org/10.1145/3432196

Brady, E., Morris, M. R., Zhong, Y., White, S., & Bigham, J. P. (2013). Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2117–2126). ACM.

Brewster, S. A., Wright, P. C., & Edwards, A. D. N. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of INTERCHI '93 Conference on Human Factors in Computing Systems* (pp. 222–227). ACM. https://doi.org/10.1145/169059.169179

Buehler, E., Branham, S., Ali, A., Chang, J. J., Hofmann, M. K., Hurst, A., & Kane, S. K. (2015). *Sharing is caring: Assistive technology designs on Thingiverse* [Paper presentation]. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15) (pp. 525–534). Association for Computing Machinery. https://doi.org/10.1145/2702123.2702525

Cao, Y., Zhang, J., Yu, Z., Liu, S., Qin, Z., Zou, Q., Du, B., & Xu, K. (2024). *CogNav: Cognitive process modeling for object goal navigation with LLMs.* arXiv preprint arXiv:2412.10439.

Carroll, J. M. (1995). *Scenario-based design: Envisioning work and technology in system development*. John Wiley & Sons.

Carroll, J. M., & Rosson, M. B. (1992). Getting around the task-artifact cycle: How to make claims and design by scenario. *ACM Transactions on Information Systems*, *10*(2), 181–212. https://doi.org/10.1145/146802.146834

Chang, R.-C., Liu, Y., & Guo, A. (2024). WorldScribe: Towards context-aware live visual descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18). ACM.

Chang, R.-C., Ting, C.-H., Hung, C.-S., Lee, W.-C., Chen, L.-J., Chao, Y.-T., Chen, B.-Y., & Guo, A. (2022). *OmniScribe: Authoring immersive audio description for 360° videos* [Paper presentation]. The 35th Annual ACM Symposium on User Interface Software and Technology (Bend, Oregon, USA) (Uist '22). Association for Computing Machinery. https://doi.org/10.1145/3526113.3545613

Cheema, M., Seifi, H., & Fazli, P. (2025). *Describe now: User-driven audio description for blind and low vision individuals* [Paper presentation]. Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25) (pp. 458–474). Association for Computing Machinery. https://doi.org/10.1145/3715336.3735685

Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Zheng, K., Lian, D., & Chen, E. (2024). When Large Language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, *27*(4), 1–45. https://doi.org/10.1007/s11280-024-01276-1

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024). YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16901–16911). IEEE.

Cockburn, A. (2000). *Writing effective use cases* (1st ed.). Addison-Wesley Longman Publishing.

Constantinescu, A., Müller, K., Haurilet, M., Petrausch, V., & Stiefelhagen, R. (2020). *Bring the environment to life: A sonification module for people with visual impairments to improve situation awareness* [Paper presentation]. Proceedings of the 2020 International Conference on Multimodal Interaction (Virtual Event, Netherlands) (ICMI '20) (pp. 50–59). Association for Computing Machinery. https://doi.org/10.1145/3382507.3418874

Constantinescu, A., Müller, K., Loitsch, C., Zappe, S., & Stiefelhagen, R. (2022). Traveling to unknown buildings: Accessibility features for indoor maps. In *Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022* (pp. 221–228). Springer-Verlag. https://doi.org/10.1007/978-3-031-08648-9_26

Constantinescu, A., Neumann, E.-M., Müller, K., Jaworek, G., & Stiefelhagen, R. (2022). Listening first: Egocentric textual descriptions of indoor spaces for people with blindness. In *Proceedings of the International Conference on Computers Helping People with Special Needs (Milan., Italy)* (pp. 241–249). Springer-Verlag. https://doi.org/10.1007/978-3-031-08648-9_28

Duh, P.-J., Sung, Y.-C., Chiang, L.-Y. F., Chang, Y.-J., & Chen, K.-W. (2020). *V-eye: A vision-based navigation system for the visually impaired* [Paper presentation]. IEEE Transactions on Multimedia 23 (pp. 1567–1580). IEEE.

Feng, J., Hamilton-Fletcher, G., Ballem, N., Batavia, M., Wang, Y., Zhong, J., Beheshti, M., Porfiri, M., & Rizzo, J.-R. (2026). Robust computer-vision based construction site detection for assistive-technology applications. *Disability and Rehabilitation. Assistive Technology*, 1–24. https://doi.org/10.1080/17483107.2026.2618130

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, *5*(1), 80–92. https://doi.org/10.1177/160940690600500107

Gamage, B., Do, T.-T., Price, N. S. C., Lowery, A., & Marriott, K. (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 (New York, NY, USA) (Assets, '23)* (pp. 30–21). Association for Computing Machinery. https://doi.org/10.1145/3597638.3608955

Gemini Team, Google. (2025). *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.* arXiv preprint arXiv:2507.06261. https://arxiv.org/pdf/2507.06261

Gonzalez Penuela, R. E., Collins, J., Bennett, C., & Azenkot, S. (2024). *Investigating use cases of AI-powered scene description applications for blind and low vision people* [Paper presentation]. CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (Chi, '24). ACM. https://doi.org/10.1145/3613904.3642211

Google. (2024). *Lookout.* https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal

Guide Dogs for the Blind Association. (2025). *Systematic search patterns.* https://www.guidedogs.org.uk/getting-support/information-and-advice/life-skills/getting-around-safely/systematic-search-patterns/.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018). VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3608–3617). IEEE.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hersh, M. (2022). Wearable travel aids for blind and partially sighted people: A review with a focus on design issues. *Sensors (Basel, Switzerland)*, *22*(14), 5454. https://doi.org/10.3390/s22145454

Herskovitz, J., Xu, A., Alharbi, R., & Guo, A. (2023). *Hacking, switching, combining: Understanding and supporting DIY assistive technology design by blind people* [Paper presentation]. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23) (pp. 57–17). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581249

Herskovitz, J., Xu, A., Alharbi, R., & Guo, A. (2024). ProgramAlly: Creating custom visual access programs via multi-modal end-user programming. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–15). ACM.

Hong, J., & Kacorri, H. (2024a). Blind users handle object recognition errors: Strategies and challenges. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (St. John's, NL, Canada) (ASSETS '24)*. Association for Computing Machinery. https://doi.org/10.1145/3663548.3675635

Hong, J., & Kacorri, H. (2024b). Understanding how blind users handle object recognition errors: Strategies and challenges. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3663548.3675635

Hong, K., He, W., Tang, H., Zhang, X., Li, Q., & Zhou, B. (2024). SPVINet: A lightweight multitask learning network for assisting visually impaired people in multiscene perception. *IEEE Internet of Things Journal*, *11*(11), 20706–20717. https://doi.org/10.1109/JIOT.2024.3371978

Hu, X., Song, A., Wei, Z., & Zeng, H. (2022). Stereopilot: A wearable target location system for blind and visually impaired using spatial audio rendering. *IEEE Transactions on Neural Systems and Rehabilitation Engineering: Publication of the IEEE Engineering in Medicine and Biology Society*, *30*(2022), 1621–1630. https://doi.org/10.1109/TNSRE.2022.3182661

Islam, M. T., Kabir, I., Pearce, E. A., Reza, M. A., & Billah, S. M. (2024). *Identifying crucial objects in blind and low-vision individuals' navigation* [Paper presentation]. Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24). ACM. https://doi.org/10.1145/3663548.3688538

Islam, R. B., Akhter, S., Iqbal, F., Saif Ur Rahman, M., & Khan, R. (2023). Deep learning based object detection and surrounding environment description for visually impaired people. *Heliyon*, *9*(6), e16924. https://doi.org/10.1016/j.heliyon.2023.e16924

Jain, G., Hindi, B., Xie, M., Zhang, Z., Srinivasula, K., Ghasemi, M., Weiner, D., Xu, X. Y. T., Paris, S. A., Tedjo, C., Bassin, J., Malcolm, M., Turkcan, M., Ghaderi, J., Kostic, Z., Zussman, G., & Smith, B. A. (2023). *Towards street camera-based outdoor navigation for blind pedestrians* [Paper presentation]. Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (New York, NY, USA) (Assets, '23). ACM. https://doi.org/10.1145/3597638.3614498

Jain, G., Teng, Y., Cho, D. H., Xing, Y., Aziz, M., & Smith, B. A. (2023). "I Want to Figure Things Out": Supporting exploration in navigation for people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction 7*(CSCW1), 1–28.

Jakob, N. (1999). *Designing web usability: The practice of simplicity*. New Riders Publishing.

Jones, B., & Kenward, M. G. (2014). *Design and analysis of cross-over trials* (3rd ed.). CRC Press.

Kacorri, H., Kitani, K. M., Bigham, J. P., & Asakawa, C. (2017). *People with visual impairment training personal object recognizers: Feasibility and challenges* [Paper presentation]. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (Denver, Colorado, USA) (CHI '17) (pp. 5839–5849). ACM. https://doi.org/10.1145/3025453.3025899

Kamikubo, R., Kato, N., Higuchi, K., Yonetani, R., & Sato, Y. (2020). *Support strategies for remote guides in assisting people with visual impairments for effective indoor navigation* [Paper presentation]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376823

Katz, B. F., Kammoun, S., Parseihian, G., Gutierrez, O., Brilhault, A., Auvray, M., Truillet, P., Denis, M., Thorpe, S., & Jouffrais, C. (2012). NAVIG: Augmented reality guidance system for the visually impaired. *Virtual Reality*, *16*(4), 253–269. https://doi.org/10.1007/s10055-012-0213-6

Kawaharazuka, K., Obinata, Y., Kanazawa, N., Tsukamoto, N., Okada, K., & Inaba, M. (2024). Reflex-based open-vocabulary navigation without prior knowledge using omnidirectional camera and multiple vision-language models. *Advanced Robotics*, *38*(18), 1307–1317. https://doi.org/10.1080/01691864.2024.2393409

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., & Lo, W.-Y. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3992–4003). IEEE.

Kolarik, A. J., Cirstea, S., Pardhan, S., & Moore, B. C. J. (2014). A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research*, *310*(2014), 60–68. https://doi.org/10.1016/j.heares.2014.01.010

KR Vision. (n.d.). *KR vision website*. http://krvision.com.cn/home

Kuribayashi, M., Ishihara, T., Sato, D., Vongkulbhisal, J., Ram, K., Kayukawa, S., Takagi, H., Morishima, S., & Asakawa, C. (2023). *PathFinder: Designing a map-less navigation system for blind people in unfamiliar buildings* [Paper presentation]. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (Chi, '23) (pp. 41–16). ACM. https://doi.org/10.1145/3544548.3580687

Kuribayashi, M., Uehara, K., Wang, A., Morishima, S., & Asakawa, C. (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 (CHI '25)*. ACM. https://doi.org/10.48550/arXiv.2502.08906

Lee, J., Herskovitz, J., Peng, Y.-H., & Guo, A. (2022). *ImageExplorer: Multi-layered touch exploration to encourage skepticism towards imperfect AI-generated image captions* [Paper presentation]. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (Chi, '22). ACM. https://doi.org/10.1145/3491102.3501966

Lee, S., Reddie, M., Tsai, C.-H., Beck, J., Rosson, M. B., & Carroll, J. M. (2020). *The emerging professional practice of remote sighted assistance for people with visual impairments* [Paper presentation]. 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20) (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376591

Lee, S., Yu, R., Xie, J., Billah, S. M., & Carroll, J. M. (2022). Opportunities for human-AI collaboration in remote sighted assistance. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (Helsinki, Finland) (IUI '22)* (pp. 63–78). Association for Computing Machinery. https://doi.org/10.1145/3490099.3511113

Li, G., Li, Z., Xia, H., & Feng, Y. (2023). Multi-sensory visual-auditory fusion of wearable navigation assistance for people with impaired vision. In *Proceedings of 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 955–960). IEEE.

Li, Y., Zheng, Y., Hamilton-Fletcher, G., Mezzavilla, M., Wang, Y., Rangan, S., Porfiri, M., Yu, Z., & Rizzo, J.-R. (2026). *Exploring the use of VLMs for navigation assistance for people with blindness and low vision.* arXiv preprint arXiv:2603.15624 (2026).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In David, F., Tomas, P., Bernt, S., & Tinne, T. (Eds.), *Computer vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740–755). Springer International Publishing.

Liu, H., Liu, R., Yang, K., Zhang, J., Peng, K., & Stiefelhagen, R. (2021). HIDA: Towards holistic indoor understanding for the visually impaired via semantic instance segmentation with a wearable solid-state LiDAR sensor. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 1780–1790). IEEE.

Liu, J. (2023). *Real-time machine learning based object detection and recognition system for the visually impaired* [Paper presentation]. 2023 Workshop on Advanced Multimedia Computing for Smart Manufacturing and Engineering (Ottawa ON, Canada) (AMC-SME '23) (pp. 31–35). ACM. https://doi.org/10.1145/3606042.3616454

Liu, J., Qiu, Z., Li, Z., Dai, Q., Zhu, J., Hu, M., Yang, M., & King, I. (2025). *A survey of personalized large language models: Progress and future directions.* arXiv abs/2502.11528 (2025). https://doi.org/10.48550/arXiv.2502.11528

Liu, R., Zhang, J., Peng, K., Zheng, J., Cao, K., Chen, Y., Yang, K., & Stiefelhagen, R. (2023). Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 1849–1859). IEEE.

Macular Disease Foundation Australia. (n.d.). *Low vision – A guide.* https://studylib.net/doc/18842187/a-guide—macular-disease-foundation-australia

Martolini, C., Cappagli, G., Saligari, E., Gori, M., & Signorini, S. (2021). Allocentric spatial perception through vision and touch in sighted and blind children. *Journal of Experimental Child Psychology*, *210*(2021), 105195. https://doi.org/10.1016/j.jecp.2021.105195

Mathis, F., & Schöning, J. (2025). LifeInsight: Design and evaluation of an AI-powered assistive wearable for blind and low vision people across multiple everyday life scenarios. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM.

Matsuki, H., Murai, R., Kelly, P. H. J., & Davison, A. J. (2024). Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18039–18048). IEEE.

Microsoft Corporation. (2024). *Seeing AI.* https://www.microsoft.com/en-us/ai/seeing-ai

Microsoft Research. (2018). *Microsoft soundscape*. https://www.microsoft.com/en-us/research/product/soundscape/

Monsaingeon, N., Caroux, L., Langlois, S., & Lemercier, C. (2023). Earcons to reduce mode confusions in partially automated vehicles: Development and application of an evaluation method. *International Journal of Human-Computer Studies*, *176*(C), 103044. https://doi.org/10.1016/j.ijhcs.2023.103044

Müller, K., Engel, C., Loitsch, C., Stiefelhagen, R., & Weber, G. (2022). Traveling more independently: A study on the diverse needs and challenges of people with visual or mobility impairments in unfamiliar indoor environments. *ACM Transactions on Accessible Computing*, *15*(2), 1–44. https://doi.org/10.1145/3514255

Natalie, R., Chang, R.-C., Sheshadri, S., Guo, A., & Hara, K. (2024). *Audio description customization* [Paper presentation]. Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (St. John's, NL, Canada) (Assets, '24) (pp. 39–19). Association for Computing Machinery. https://doi.org/10.1145/3663548.3675617

netz-barrierefrei.de. (n.d.). *What is blind life?* https://netz-barrierefrei.de/en/what-is-blind-life.html

OpenAI. (2023). *Prompt engineering guide.* https://platform.openai.com/docs/guides/prompt-engineering

Ottink, L., van Raalte, B., Doeller, C. F., Van der Geest, T. M., & Van Wezel, R. J. A. (2022). Cognitive map formation through tactile map navigation in visually impaired and sighted persons. *Scientific Reports*, *12*(1), 11567. https://doi.org/10.1038/s41598-022-15858-4

Ou, W., Zhang, J., Peng, K., Yang, K., Jaworek, G., Müller, K., & Stiefelhagen, R. (2022). Indoor navigation assistance for visually impaired people via dynamic SLAM and panoptic segmentation with an RGB-D sensor. In *Proceedings of International Conference on Computers Helping People with Special Needs* (pp. 160–168). Springer. https://doi.org/10.1007/978-3-031-08648-9_19

Oumard, C., Kreimeier, J., & Götzelmann, T. (2022). *Implementation and evaluation of a voice user interface with offline speech processing for people who are blind or visually impaired* [Paper presentation]. The 15th International Conference on PErvasive Technologies Related to Assistive Environments (Corfu, Greece) (PETRA '22) (pp. 277–285). ACM. https://doi.org/10.1145/3529190.3529197

Perkins School for the Blind. (2024). *How to write alt text and image descriptions for the visually impaired*. https://www.perkins.org/resource/how-write-alt-text-and-image-descriptions-visually-impaired/

Piccinelli, L., Yang, Y.-H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., & Yu, F. (2024). Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10106–10116). IEEE.

Prem, S. (1995). *Mann introductory statistics* (2nd ed.). Wiley.

Reginald, G. (Ed.). (1999). *Wayfinding behavior: Cognitive mapping and other spatial processes*. Johns Hopkins University Press.

Reimers, N., & Gurevych, I. (2021). *all-MiniLM-L6-v2*. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Riddering, A. (2023). *Scanning efficiently for activities of daily living*. VisionAware. https://aphconnectcenter.org/visionaware/eye-conditions/eye-health/low-vision/scanning-efficiently-for-activities-of-daily-living/

Russell, D. M., & Chi, E. H. (2014). Looking back: Retrospective study methods for HCI. In Judith S. O., & Wendy A., K. (Eds.), *Ways of knowing in HCI* (pp. 373–393). Springer. https://doi.org/10.1007/978-1-4939-0378-8_15

Schauerte, B., Martinez, M., Constantinescu, A., & Stiefelhagen, R. (2012). An assistive vision system for the blind that helps find lost things. In *Proceedings of the 13th International Conference Computers Helping People with Special Needs* (pp. 566–572). Springer. https://doi.org/10.1007/978-3-642-31534-3_83

Schinazi, V. R., Thrash, T., & Chebat, D.-R. (2016). Spatial navigation by congenitally blind individuals. *Wiley Interdisciplinary Reviews. Cognitive Science*, *7*(1), 37–58. https://doi.org/10.1002/wcs.1375

Scott, B. (2025). *Orientation & mobility*. https://glaucoma.org.au/orientation-and-mobility. Glaucoma Australia

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, Z., Ebrahimi, S., & Wang, H. (2026). Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, *58*(5), 1–42. https://doi.org/10.1145/3735633

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)* (pp. 336). IEEE Computer Society.

Singh Chaplot, D., Gandhi, D. P., Gupta, A., & Salakhutdinov, R. R. (2020). Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, *33*, 4247–4258.

Sugashini, T., & Balakrishnan, G. (2024). YOLO glass: Video-based smart object detection using squeeze and attention YOLO network. *Signal, Image and Video Processing*, *18*(3), 2105–2115. https://doi.org/10.1007/s11760-023-02855-x

Sun, J., Wu, J., Ji, Z., & Lai, Y.-K. (2025). A survey of object goal navigation. *IEEE Transactions on Automation Science and Engineering*, *22*(2025), 2292–2308. https://doi.org/10.1109/TASE.2024.3378010

Supersense. (2024). *Supersense*. https://www.supersense.app/

Surougi, H. R., & McCann, J. A. (2023). Real-time optimisation-based path planning for visually impaired people in dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 1831–1840). IEEE.

Tahoun, N., Awad, A., & Bonny, T. (2020). *Smart assistant for blind and visually impaired people* [Paper presentation]. Proceedings of the International Conference on Advances in Artificial Intelligence (Istanbul, Turkey) (ICAAI '19) (pp. 227–231). ACM. https://doi.org/10.1145/3369114.3369139

Taioli, F., Zorzi, E., Franchi, G., Castellini, A., Farinelli, A., Cristani, M., & Wang, Y. (2024). *Collaborative instance navigation: Leveraging agent self-dialogue to minimize user input*. arXiv preprint arXiv:2412.01250.

Tan, Z., Liu, Z., & Jiang, M. (2024). *Personalized pieces: Efficient personalized large language models through collaborative efforts* [Paper presentation]. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 6459–6475). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.371

TapTapSee. (n.d). *TapTapSee*. https://www.taptapseeapp.com Version 3.1.1 [Mobile application software].

Turkstra, L. M., Bhatia, T., Van Os, A., & Beyeler, M. (2025). Assistive technology use in domestic activities by people who are blind. *Scientific Reports*, *15*(1), 7486. https://doi.org/10.1038/s41598-025-91755-w

V7. (2024). *Aipoly*. https://www.aipoly.com

Varghese, R., & Sambath, M. (2024). *YOLOv8: A novel object detection algorithm with enhanced performance and robustness* [Paper presentation]. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS) (pp. 1–6). ACM. https://doi.org/10.1109/ADICS58448.2024.10533619

Waisberg, E., Ong, J., Masalkhi, M., Zaman, N., Sarker, P., Lee, A. G., & Tavakkoli, A. (2024). Meta smart glasses—Large language models and the future for assistive glasses for individuals with vision impairments. *Eye (London, England)*, *38*(6), 1036–1038. https://doi.org/10.1038/s41433-023-02842-z

Wang, H., Qin, J., Bastola, A., Chen, X., Suchanek, J., Zihao, G., & Abolfazl, R. (2024). *VisionGPT: LLM-assisted real-time anomaly detection for safe visual navigation.* arXiv Preprint, arXiv:2403.12415.

Wen, L. Y., Morrison, C., Grayson, M., Marques, R. F., Massiceti, D., Longden, C., & Cutrell, E. (2024). *Find my things: Personalized accessibility through teachable AI for people who are blind or low vision* [Paper presentation]. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery. https://doi.org/10.1145/3613905.3648641

Withagen, A., Kappers, A. M. L., Vervloed, M. P. J., Knoors, H., & Verhoeven, L. (2013). The use of exploratory procedures by blind and sighted adults and children. *Attention, Perception & Psychophysics*, *75*(7), 1451–1464. https://doi.org/10.3758/s13414-013-0479-0

World Health Organization. (2016). *Priority assistive products list.* https://www.who.int/publications/i/item/priority-assistive-products-list

World Health Organization. (2021). *ICD-10: International statistical classification of diseases and related health problems, 10th Revision.* https://icd.who.int/browse10/2021/en/H54. Accessed: 2023-08-05.

Woźniak, S., Koptyra, B., Janz, A., Kazienko, P., & Kocoń, J. (2024). *Personalized large language models*. arXiv abs/2402.09269. https://doi.org/10.48550/arXiv.2402.09269

Xie, J., Reddie, M., Lee, S., Billah, S. M., Zhou, Z., Tsai, C.-H., & Carroll, J. M. (2022). Iterative design and prototyping of computer vision mediated remote sighted assistance. *ACM Transactions on Computer-Human Interaction: Publication of the Association for Computing Machinery*, *29*(4), 1–40. https://doi.org/10.1145/3501298

Xie, J., Yu, R., Lee, S., Lyu, Y., Billah, S. M., & Carroll, J. M. (2022). Helping helpers: Supporting volunteers in remote sighted assistance with augmented reality maps. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (Virtual Event, Australia) (DIS '22)* (pp. 881–897). Association for Computing Machinery. https://doi.org/10.1145/3532106.3533560

Xie, J., Yu, R., Zhang, H., Billah, S. M., Lee, S., & Carroll, J. M. (2025). Beyond visual perception: Insights from smartphone interaction of visually impaired users with large multimodal models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM.

Xie, J., Yu, R., Zhang, H., Lee, S., Billah, S. M., & Carroll, J. M. (2024). Emerging practices for large multimodal model (LMM) assistance for people with visual impairments: Implications for design. *arXiv preprint arXiv: 2407.08882*.

Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., & Zhang, Y. (2025). *A-mem: Agentic memory for LLM agents*. arXiv preprint arXiv:2502.12110.

Yan, C., Qu, D., Xu, D., Zhao, B., Wang, Z., Wang, D., & Li, X. (2024). GS-SLAM: Dense visual slam with 3D Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19595–19604). IEEE.

Yang, B., He, L., Liu, K., & Yan, Z. (2024). *VIAssist: Adapting multi-modal large language models for users with visual impairments* [Paper presentation]. Proceedings of the IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys) (pp. 32–37). IEEE. https://doi.org/10.1109/FMSys62467.2024.00010

Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L., & Xie, S. (2025). Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 10632–10643). IEEE.

Yang, K., Bergasa, L. M., Romera, E., Cheng, R., Chen, T., & Wang, K. (2018). Unifying terrain awareness through real-time semantic segmentation. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)* (pp. 1033–1038). IEEE.

Yang, K., Bergasa, L. M., Romera, E., Huang, X., & Wang, K. (2018). Predicting polarization beyond semantics for wearable robotics. In *Proceedings of the IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)* (pp. 96–103). IEEE.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything v2. *Advances in Neural Information Processing Systems*, *37*, 21875–21911. https://proceedings.neurips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html

Yang, X., Tian, Y., Yi, C., & Arditi, A. (2010). *Context-based indoor object detection as an aid to blind persons accessing unfamiliar environments* [Paper presentation]. Proceedings of the 18th ACM International Conference on Multimedia (Firenze, Italy) (MM '10) (pp. 1087–1090). Association for Computing Machinery. https://doi.org/10.1145/1873951.1874156

Yang, Y., Yang, H., Zhou, J., Chen, P., Zhang, H., Du, Y., & Gan, C. (2025). 3D-mem: 3D scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 17294–17303). IEEE.

Yi, C., Flores, R. W., Chincha, R., & Tian, Y. (2013). Finding objects for assisting blind people. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, *2*(2), 71–79. https://doi.org/10.1007/s13721-013-0026-x

Yin, H., Xu, X., Wu, Z., Zhou, J., & Lu, J. (2024). *SG-Nav: Online 3D scene graph prompting for LLM-based zero-shot object navigation* [Paper presentation]. Advances in Neural Information Processing Systems. IEEE.

Yin, H., Xu, X., Zhao, L., Wang, Z., Zhou, J., & Lu, J. (2025). UniGoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE.

Yokoyama, N., Ramrakhya, R., Das, A., Batra, D., & Ha, S. (2024). *HM3D-OVON: A dataset and benchmark for open-vocabulary object goal navigation* [Paper presentation]. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 5543–5550). IEEE.

Zhang, D., Qiao, J. U., Kim, S.-H., Bae, S., Lee, C., & S., Hong. (2023). *Han, Yu faster segment anything: Towards lightweight SAM for mobile applications.* arXiv preprint arXiv:2306.14289.

Zhang, H., Falletta, N. J., Xie, J., Yu, R., Lee, S., Billah, S. M., & Carroll, J. M. (2025). *Enhancing the travel experience for people with visual impairments through multimodal interaction: NaviGPT, A real-time AI-driven mobile navigation system* [Paper presentation]. Companion Proceedings of the 2025 ACM International Conference on Supporting Group Work (Hilton Head, New Jersey, USA) (GROUP '25) (pp. 29–35). Association for Computing Machinery. https://doi.org/10.1145/3688828.3699636

Zhang, J., Yang, K., Constantinescu, A., Peng, K., Müller, K., & Stiefelhagen, R. (2021). Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 1760–1770). IEEE.

Zhao, Y., Zhang, Y., Xiang, R., Li, J., & Li, H. (2024). *VIALM: A survey and benchmark of visually impaired assistance with large models.* arXiv preprint arXiv:2402.01735.

Zheng, J., Zhang, J., Yang, K., Peng, K., & Stiefelhagen, R. (2024). MateRobot: Material recognition in wearable robotics for people with visual impairments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2303–2309). IEEE.

Zhu, S., Wang, G., Blum, H., Liu, J., Song, L., Pollefeys, M., & Wang, H. (2024). SNI-SLAM: Semantic neural implicit slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21167–21177). IEEE.

Zou, W., Hua, G., Zhuang, Y., & Tian, S. (2023). Real-time passable area segmentation with consumer RGB-D cameras for the visually impaired. *IEEE Transactions on Instrumentation and Measurement*, *72*, Article 2513011. https://doi.org/10.1109/TIM.2023.3272369

About the authors

Ruiping Liu is a PhD student at the Computer Vision for Human–Computer Interaction Lab (cv:hci), Karlsruhe Institute of Technology (KIT), Germany. Her research focuses on assistive technology for people with visual impairments, embodied AI, and computer vision.

Jiaming Zhang is Professor at the School of AI and Robotics, Hunan University (HNU), China. He received his PhD (summa cum laude) from Karlsruhe Institute of Technology (KIT), where he remains a collaborating researcher. His research focuses on computer vision, scene understanding, and assistive systems for people with visual impairments.

Angela Schön is a researcher at the Center for Digital Accessibility and Assistive Technology (ACCESS@KIT), Karlsruhe Institute of Technology (KIT), Germany. She received her PhD from KIT on auditory displays for people with visual impairments during travel. Her research focuses on acoustic interfaces and accessible navigation.

Karin Müller is Deputy Director of the Center for Digital Accessibility and Assistive Technology (ACCESS@KIT), Karlsruhe Institute of Technology (KIT), Germany, and co-leads the Real-World Lab Accessibility. Her work centers on accessible mobility, orientation training, and assistive technologies for people with visual impairments.

Junwei Zheng is a PhD student at the Computer Vision for Human–Computer Interaction Lab (cv:hci), Karlsruhe Institute of Technology (KIT), Germany, and a visiting PhD student at ETH Zurich. His research focuses on scene understanding, visual localization, and vision-language navigation for people with visual impairments.

Kailun Yang received his PhD degree in Information Sensing and Instrumentation from the State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang University (ZJU) in 2019. He is currently a professor at the School of Artificial Intelligence and Robotics, Hunan University (HNU).

Anhong Guo is the Morris Wellman Assistant Professor in Computer Science and Engineering at the University of Michigan. He received his PhD from Carnegie Mellon University. His research focuses on human–AI interaction and accessibility. He is a recipient of the NSF CAREER award, and a Forbes 30 Under 30 scientist.

Kathrin Gerling is Professor of Human–Computer Interaction and Accessibility at Karlsruhe Institute of Technology (KIT), Germany, where she leads the HCI and Accessibility research group (https://hci.iar.kit.edu). Her research focuses on how interactive technology can be designed in a way that supports human self-determination.

Rainer Stiefelhagen is Professor and head of the Computer Vision for Human–Computer Interaction Lab (cv:hci) and ACCESS@KIT at Karlsruhe Institute of Technology (KIT), Germany, and co-director of the Real-World Lab Accessibility. His research spans computer vision, multimodal human–computer interaction, and assistive systems for people with visual impairments.